STRUCTURED KINEMATIC ROUTING FOR UNIFIED ACTION DIFFERENCE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Analyzing fine-grained execution differences in skilled activities, such as sports, demands moving beyond holistic feature extraction to precise, domain-specific visual reasoning. While prior work has made progress in evaluating individual performance, existing methods fall short in comparing two similar actions (e.g., penalty kick in soccer) conducted by different performers and explaining *how* their actions differ. To address this gap, we introduce **Unified Action Difference Reasoning (UADR)**, a novel task that unifies *quantitative* performance scores and *qualitative* explanations of inter-performer differences, enabling actionable feedback for improvement. To support this task, we construct the ADR dataset, built upon Ego-Exo4D dataset, comprising paired videos annotated with both performance scores and natural language descriptions of action differences. We further propose ***S**tructured **K**inematic **R**outing (**SKR**)*, a differentiable kinematic routing framework that models the visual reasoning process behind performance differences by capturing fine-grained kinematic cues. SKR discovers an optimal visual trajectory through a hierarchical kinematic ontology, regularizes these paths via cross-modal semantic grounding using expert commentary priors, and captures kinematic topological deviations through kinematic discrepancy reasoning. Comprehensive experiments demonstrate that SKR significantly outperforms state-of-the-art vision-language models and baseline approaches on the UADR task. Furthermore, our framework generalizes effectively to traditional Action Quality Assessment (AQA) settings, surpassing state-of-the-art approaches on benchmarks including T3SeT and FitnessAQA. Code, model and dataset will be released upon the completion of review process.

1 INTRODUCTION

Analyzing and comparing skilled activities (e.g. sports) is highly valuable for education and coaching, supporting trainees across different skill levels. For novice athletes, like beginner soccer players, precise analysis of actions (e.g., ball control, kicking) helps decompose complex movements into fundamental components. For advanced performers, fine-grained action analysis enables targeted feedback for skill refinement and performance optimization. Crucially, effective training requires both *quantitative* assessment and *qualitative* interpretation of fine-grained actions. However, analyzing actions in isolation is often insufficient. A more effective approach involves comparing the same actions across different performers to uncover subtle execution differences. These detailed comparisons reveal specific areas for improvement, providing actionable insights that enable performers to refine their techniques and accelerate skill development.

Prior work on analyzing skilled activities Xu et al. (2024c); Zhou et al. (2024b); Majeedi et al. (2024); Bai et al. (2022); Yun et al. (2024); Xu et al. (2024a; 2022b) primarily evaluates individual performers in isolation. While effective for assessing overall performance quality, such methods are limited in their ability to capture and explain fine-grained technical differences across performers. This restriction hinders comparative analysis, which is essential for revealing subtle execution variations that can inform personalized feedback and targeted skill refinement. Burgess et al. Burgess et al. (2025) recently introduced the video action differencing task to compare how two individuals perform the same actions. However, the method provides only qualitative analysis without quantitative evaluation. Other approaches, such as ExpertAF Ashutosh et al. (2025) and CrossTrainer Ashutosh &

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

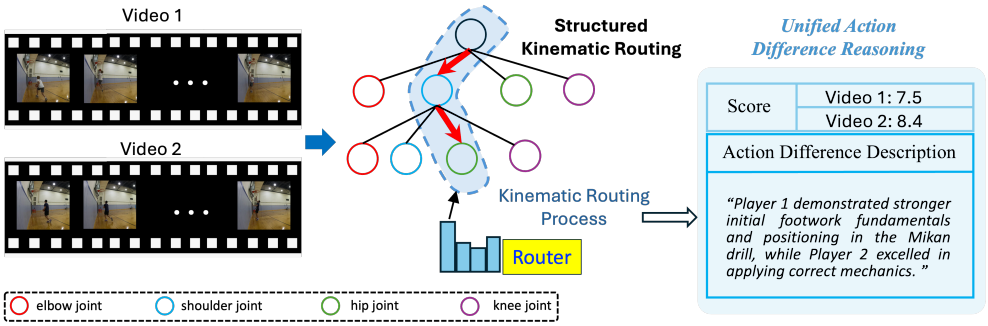

Figure 1: Overview of the unified action difference reasoning (UADR) task and our proposed Structured Kinematics Routing approach. Our model takes two videos doing the similar action as inputs, generating both *quantitative* score predictions and *qualitative* action difference descriptions between the two videos.

Grauman (2025), offer actionable feedback but still lack accurate quantitative assessment, which is crucial for performers to objectively understand and track their performance level.

We address this gap by introducing the new **Unified Action Difference Reasoning (UADR)** task. As shown in Figure 1, given two videos of performers executing the same action, UADR requires a model to output two performance scores (*quantitative*) and a natural-language description of their action differences (*qualitative*). This task is challenging because the model must simultaneously assess individual action quality and reason about pairwise discrepancies. The relevant differences are often subtle and localized: small variations in joint angles, movement velocity, body posture, or coordination between body parts can lead to large differences in perceived performance quality. Capturing such cues requires moving beyond global video-level representations toward structured reasoning over human kinematics.

To this end, we propose the ***Structured Kinematic Routing (SKR)*** framework. The key idea is to formulate comparative action assessment as routing through a structured space of anatomy-grounded kinematic primitives. Instead of directly regressing scores or generating descriptions from visual features, SKR first constructs a kinematic-routing tree over major body joints and their physical motion attributes. Each path in the tree represents a candidate assessment route, corresponding to an ordered inspection of body regions and kinematic states. The model selects video-conditioned routes that capture the most relevant evidence for evaluating performance.

Our SKR framework is built around three major components. The first is an *Anatomy-grounded Kinematic Reasoning Tree*, which organizes joint-level motion attributes such as angle, velocity, and 3D position into structured candidate routes. The second is *Cross-Modal Semantic Grounding*, which aligns selected kinematic routes with expert commentary during training, encouraging the latent reasoning process to reflect semantically meaningful assessment cues. The third is *Kinematic Discrepancy Reasoning*, which constructs implicit kinematic relation graphs for each performer and converts their residual differences into graph tokens for score prediction and action-difference generation. Through these components, SKR provides a shared structured representation for both quantitative and qualitative reasoning.

We further construct ADR, a new benchmark for UADR based on the Ego-Exo4D Grauman et al. (2024) dataset. ADR contains paired videos from basketball, soccer, and rock climbing, with expert-derived performance scores and natural-language action-difference descriptions. This benchmark allows us to evaluate whether models can jointly predict performance quality and explain inter-performer differences in skilled activities.

Experiments show that SKR consistently outperforms action quality assessment models and large vision-language baselines on the ADR dataset. It improves both score prediction and action-difference description generation, indicating that structured kinematic routing helps models capture fine-grained performance-relevant cues. SKR also generalizes to standard action quality assessment (AQA) datasets, demonstrating that anatomy-grounded routing and residual kinematic discrepancy modeling are broadly useful for fine-grained human action understanding. Our contributions of this paper are summarized as follows:

(1) We introduce Unified Action Difference Reasoning (UADR), a task that jointly predicts individual action quality scores and generates natural-language explanations of differences between paired performances.
(2) We propose Structured Kinematic Routing (SKR), a framework that models action assessment through an anatomy-grounded kinematic routing tree, cross-modal semantic grounding, and residual kinematic discrepancy reasoning.
(3) We construct the ADR dataset, which includes both score-based evaluations and expert-generated comments on action differences. Comprehensive experiments demonstrate that our approach significantly outperforms existing state-of-the-art methods on both this challenging UADR task and traditional AQA task, highlighting its effectiveness in detailed sports performance analysis.

2 Action Difference Reasoning Dataset

To support research on unified action difference reasoning, we constructed a new dataset, *ADR*, derived from the Ego-Exo4D dataset Grauman et al. (2024). *ADR* includes videos from three sport categories: soccer, basketball, and rock climbing. Soccer and basketball activities are divided into three sub-activities each, while rock climbing is represented by a single sub-activity. As detailed in Table 4 in appendix, the basketball activities include Mid-Range Jump Shooting, Mikan Layup, and Reverse Layup, while soccer activities are categorized as Penalty Kick, Juggling, and Dribbling. To maintain clean scenarios, we selected one exocentric video per player from the original dataset, choosing only videos where no other unrelated individuals are present. More analysis on designing the dataset can be found in Sec C of appendix.

Since every player is commented by several experts with individual score evaluations in the original Ego-Exo4D dataset, the ground-truth score for each player is obtained by averaging all the scores provided by experts. Expert comment for each video is also obtained by combining all expert comments in the Ego-Exo4D dataset. To construct the *ADR* dataset, we group videos in each sub-activity into video pairs, as the examples shown in Figure 2. Then the ground-truth action difference descriptions are generated by prompting Gemini 2.5 Google DeepMind (2025) with expert comments on each video provided from the Ego-Exo4D dataset. Figure 2 shows the process of constructing ground-truth action difference descriptions. To ensure that the generated descriptions focus on action differences of players, the instruction prompt for Gemini 2.5 is designed to stress that the summary should include which part does which player do better, as shown in Figure 2. In total, our curated *ADR* dataset contains over 3,700 ego–exo video pairs spanning three distinct sports categories. The number of video pairs generated for each sports category is also listed in Table 4 in appendix.

To validate the quality of action difference descriptions generated by Gemini 2.5, we conducted a systematic human evaluation. Details on human evaluation are presented in Sec C of appendix.

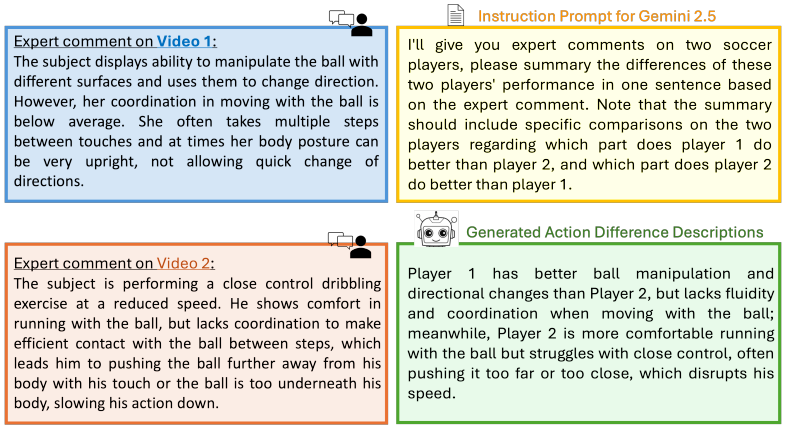

Figure 2: The process of data construction for action difference description generation with Gemini 2.5. We take the instruct prompt and expert comments on video 1 and video 2 as inputs to Gemini 2.5, generating action difference descriptions.

3 Structured Kinematic Routing

Human evaluation of complex, skilled activities relies on structured, hierarchical attention. Experts do not passively process global visual features. Instead, they navigate an internal anatomical ontology, selectively shifting their diagnostic gaze to identify critical kinematic deviations.

To model this expert cognitive process, we introduce the ***Structured Kinematic Routing (SKR)*** framework. As illustrated in Figure 3, SKR comprises three core mechanisms: **(1) Anatomy-grounded Kinematic Routing Tree**, which formalizes the sequence of visual fixations through a combinatorial search space; **(2) Cross-Modal Semantic Grounding**, which regularizes these learned visual trajectories using expert commentary priors; and **(3) Kinematic Discrepancy Reasoning**, which captures kinematic topological deviations to articulate exact performance differences.

This design provides two advantages. First, it introduces an interpretable latent structure between raw video perception and final assessment, allowing the model to expose which body regions are most relevant during evaluation. Second, it enables pairwise comparison at the level of residual kinematic relations, which is crucial for describing subtle technical differences between performers.

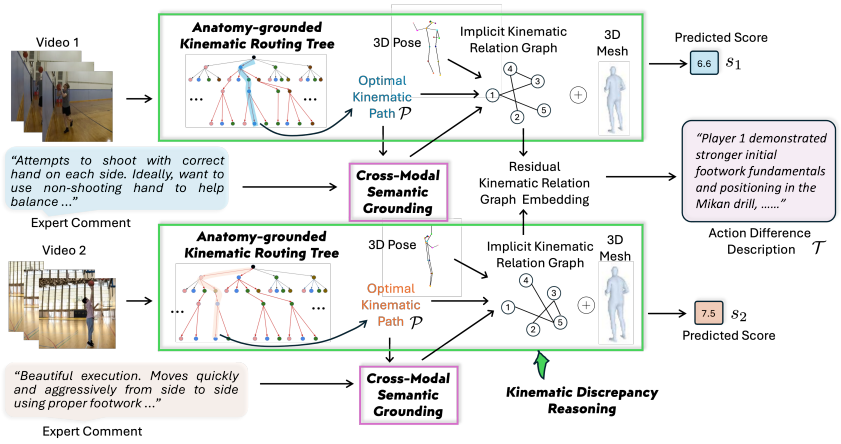

Figure 3: Overall pipeline of the SKR framework, including *Anatomy-grounded Kinematic Reasoning Tree*, *Cross-Modal Sematic Grounding* and *Kinematic Discrepancy Reasoning* modules.

3.1 Task Formulation

We introduce the novel task called Unified Action Difference Reasoning (UADR) that unifies both quantitative score assessment and qualitative language descriptions on action differences between two performers doing the same action. Specifically, given two videos V_1 and V_2 that doing the same category of action a, UADR task requires to output both quantitative results and qualitative results: (1) predicted scores s_1 and s_2 of the two corresponding videos (*Quantitative*); (2) language descriptions $\mathcal{T}$ of the difference of the two videos (*Qualitative*).

Our objective is to learn a scorer g_θ parametrized by θ that generates these two components: $s_1 = g_\theta(V_1), s_2 = g_\theta(V_2)$ and an action difference descriptor h_θ to generate the language descriptions $\mathcal{T} = h_\theta(V_1, V_2)$. During training, each video is associated with an expert score, and each video pair is associated with an action-difference description. During inference, only the two videos are provided. The model must infer both individual performance quality and comparative technical differences.

3.2 Anatomy-Grounded Kinematic Routing Tree

A key challenge in action difference reasoning is that the relevant evidence is often localized and sequential. For example, when evaluating a basketball layup, an expert may first inspect footwork, then body balance, and finally hand placement near release. Directly pooling global video features can obscure such local and temporal cues. To address this, we formulate action assessment as a structured routing problem over anatomical motion primitives.

Kinematic state construction. As shown in Figure 4, we define a compact set of anatomical primitives corresponding to major movement-related joints: *elbow*, *shoulder*, *hip*, and *knee*. Each node in the kinematic routing tree represents one anatomical primitive at a particular reasoning step. For a given joint, we encode its local kinematic state using three types of physical information: $m = (a, v, p)$, where a denotes the joint angle, v denotes joint velocity, and $p = (x, y, z)$ denotes the normalized 3D joint position. These physical statistics provide an explicit kinematic representation of body motion and serve as the basic units of the kinematic routing tree. The tree is constructed with finite depth K. A root-to-leaf path therefore defines a reasoning trajectory $l = (m_1, m_2, \ldots, m_K)$, where each m_k is a kinematic state associated with a selected anatomical primitive. This formulation turns expert-style action assessment into a structured latent path-selection problem.

Visual reasoning routing scoring. Given a video V, we score each candidate reasoning trajectory according to how well it explains the action-quality evidence in the video. Since routing vocabulary has four primitive choices per hop, the full kinematic routing tree contains $4^4 = 256$ candidate paths (trajectories) per video. Let $f_V = E_V(V) \in \mathbb{R}^{N\times d}$ be the video feature extracted by a pretrained video encoder E_V, where d indicates the feature dimension. Let $f_M \in \mathbb{R}^{N\times d}$ denote the sequence of kinematic statistics associated with a candidate path. We project both visual and kinematic representations into a shared feature space: $f'_M = \mathcal{F}_M(f_M) \in \mathbb{R}^{N\times d}, \qquad f'_V = \mathcal{F}_V(f_V) \in \mathbb{R}^{N\times d}$, where N is the number of video keyframes in a video input. The selected reasoning trajectory l^* is obtained by selection the path with highest score:

$$Q(l) = \sigma\left(\mathcal{F}_R\left(\text{MHCA}(f'_M, f'_V)\right)\right), \tag{1}$$

where $Q(l) \in [0, 1]$ is the trajectory score, $\mathcal{F}_M$, $\mathcal{F}_V$, and $\mathcal{F}_R$ are learnable projection networks, σ is the Sigmoid function, MHCA indicates multi-head cross attention module in the Transformer block. This trajectory-selection mechanism can be viewed as searching over a discrete space of latent assessment programs. Each program specifies which body regions should be inspected and in what order.

Cross-video routing consistency. For two videos belonging to the same action category, the model should generally rely on comparable assessment logic, even though the performers may differ in quality. To encourage this behavior, we regularize the route-score distributions of paired videos. Let $p_r(l_i) = \frac{Q(l_i)}{\sum_j Q(l_j)}$ denote the normalized probability for trajectory l_i. For a pair of videos V_1 and V_2, we minimize:

$$\mathcal{L}_{KL} = \text{KL}(p_r^1 \| p_r^2), \tag{2}$$

where p_r^1 and p_r^2 are routing distributions of the two videos. This loss encourages the model to compare performers through a shared kinematic reasoning basis, rather than relying on unrelated video cues.

3.3 Cross-Modal Semantic Grounding

The anatomy-grounded routing tree defines a structured search space, but not every selected route is necessarily semantically meaningful. To make the learned routing process better aligned with expert assessment, we ground the selected kinematic route in both visual evidence and expert language. The key idea is to use expert commentary as training-time semantic supervision. Expert comments describe what aspects of the movement matter for action quality, such as balance, coordination, control, footwork, or body alignment. By aligning selected kinematic routes with expert commentary, SKR learns routes that are not only visually predictive but also semantically consistent with human assessment.

Let $f_C \in \mathbb{R}^d$ denote the language embedding extracted from expert commentary using a language encoder. Given the selected kinematic path representation $f'_M \in \mathbb{R}^d$, we compute two cross-attended representations: $f_{CM} = \text{MHCA}(f_C, f'_M) \in \mathbb{R}^{N\times d}$, $f_{VM} = \text{MHCA}(f_V, f'_M) \in \mathbb{R}^{N\times d}$, where f_C and f_V serve as query, f_M serves as key and value in MHCA. We align these two representations using a semantic consistency loss:

$$\mathcal{L}_{semantic} = 1 - \cos\left(\bar{f}_{CM}, \bar{f}_{VM}\right), \tag{3}$$

where $\bar{f}_{CM} \in \mathbb{R}^d$ and $\bar{f}_{VM} \in \mathbb{R}^d$ are average-pooled embeddings from $f_{CM} \in \mathbb{R}^{N\times d}$ and $f_{VM} \in \mathbb{R}^{N\times d}$. This alignment objective encourages the latent reasoning trajectory to capture the same action-relevant cues that experts describe in language. Importantly, expert commentary is only required during training. At inference time, the model performs reasoning directly from the input videos.

After semantic grounding, the video-grounded route representation f_{VM} is used to enhance keypoint and mesh features. Let $\mathbf{K} = E_{KP}(V) \in \mathbb{R}^{P\times 3}$ denote the extracted 3D keypoints from an existing keypoint extractor E_{KP}, where P is the number of keypoints. We first map them into a keypoint feature space: $f_K = \mathcal{F}_K(\mathbf{K}) \in \mathbb{R}^{P\times d}$. We then update the keypoint features using the grounded route representation: $f'_K = \text{MHCA}(f_{VM}, f_K)$. The resulting f'_K is a knowledge-enhanced keypoint representation, emphasizing body joints that are relevant to the selected kinematic route. To incorporate richer 3D body geometry, we also extract a human mesh $\mathbf{M} = E_{ME}(V)$ with a 3D mesh extractor E_{ME}. We further map it into mesh features $f_{ME} = \mathcal{F}_{ME}(\mathbf{M}) \in \mathbb{R}^{P\times L}$, where L denotes the dimension of mesh embeddings. The mesh representation is similarly enhanced by the

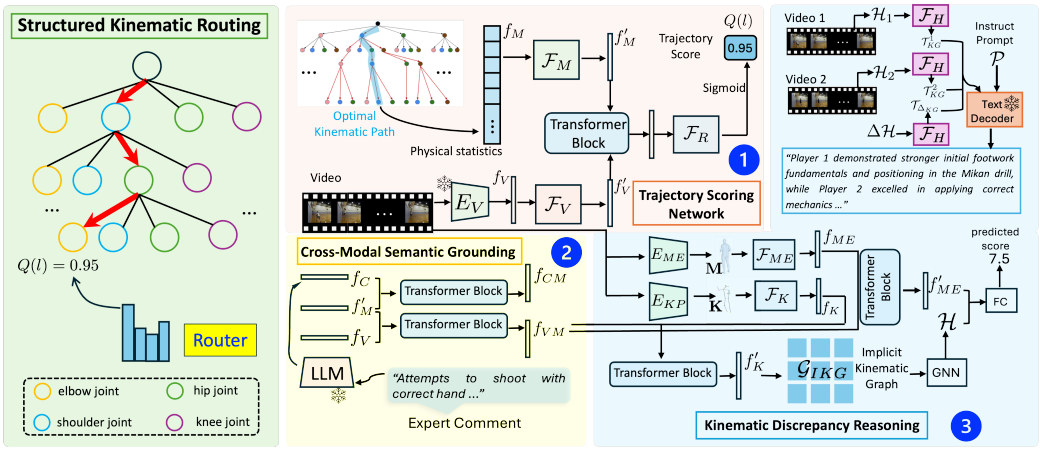

Figure 4: Detailed pipeline of unified action difference reasoning. **(Left)** Structured Kinematic Routing. **(Right)** Three major components in the pipeline: ① Trajectory Scoring Network based on anatomy-grounded kinematic routing tree; ② Cross-Modal Semantic Grounding; ③ Kinematic Discrepancy Reasoning. Each module is highlighted with different color. Detailed illustrations can be found in Section 3.

grounded route representation: $f'_{ME} = \text{MHCA}(f_{VM}, f_{ME}) \in \mathbb{R}^{P\times L}$. Thus, cross-modal grounding transforms raw keypoint and mesh features into route-aware kinematic representations. These representations serve as the basis for both score prediction and action-difference reasoning.

3.4 Kinematic Discrepancy Reasoning

The final stage uses the grounded kinematic representations to reason about action quality and inter-performer differences. This stage has two goals: predicting individual action-quality scores and generating an action-difference description for the video pair.

Implicit kinematic relation graph. Action quality depends not only on individual joints but also on dependencies between body parts. For instance, poor movement quality may result from weak coordination between the shoulder and elbow, or from misalignment between the hip and knee. To model such dependencies, we construct an implicit kinematic relation graph over the knowledge-enhanced keypoint features. We start from the physical skeleton graph $\mathcal{G} \in \mathbb{R}^{P\times P}$, where $G_{ij} = 1$, if keypoints i and j are anatomically connected, and $G_{ij} = 0$ otherwise. This graph captures explicit physical connections in the human skeleton. However, action quality may also depend on non-adjacent joint interactions. Therefore, we construct an implicit kinematic graph $\mathcal{G}_{IKG}$ over non-adjacent keypoint pairs. Given two enhanced keypoint embeddings $x_i \in \mathbb{R}^d$ and $x_j \in \mathbb{R}^d$, we estimate their dependency probability as

$$P(x_i, x_j|\theta) = \frac{1}{Z(\theta)}\exp\big(-\phi(x_i, x_j; \theta)\big), \quad Z(\theta) \triangleq \sum_{(x,x')\in\mathcal{P}} \exp\big(-\phi(x, x'; \theta)\big), \tag{4}$$

where $\phi(\cdot)$ is a learnable pairwise function and $Z(\theta)$ is the normalization factor. We threshold these probabilities to obtain the implicit kinematic relation graph $\mathcal{G}_{IKG}$.

Graph reasoning for score prediction. Given the explicit skeleton graph $\mathcal{G}$, the implicit kinematic relation graph $\mathcal{G}_{IKG}$, and the enhanced keypoint features f'_K, we compute the final kinematic graph embedding: $\mathcal{H} \in \mathbb{R}^{P\times d}$: $\mathcal{H} = \sigma\big(\mathcal{G}_{\text{IKG}}\big(\sigma(\mathcal{G} f'_K \mathbf{W})\big)\mathbf{W}_{\text{IKG}}\big)$, where σ indicates the nonlinear activation function, $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_{\text{IKG}} \in \mathbb{R}^{d\times d}$ are learnable weights. The predicted score $\hat{s}$ for each video segment is generated with a light-weight fully connected layer $\mathbf{FC}$: $\hat{s} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{FC}(\mathcal{H}, f'_{ME})$, where K denotes the number of segments for the entire video. The score prediction process is supervised with the score loss function $\mathcal{L}_{\text{score}}$:

$$\mathcal{L}_{\text{score}} = ||\hat{s} - s||_2^2, \tag{5}$$

where s denotes the ground-truth score of the video clip. $\mathcal{L}_{\text{score}}$ aims to minimize the difference between the model-predicted action scores and the ground-truth values, ensuring that the model accurately evaluates action quality.

Residual kinematic discrepancy tokens. While score prediction evaluates each video individually, action-difference generation requires explicit comparison between the two performers. We therefore introduce residual kinematic discrepancy tokens. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ denote the graph embeddings of videos V_1 and V_2. We first map each graph embedding into keypoint graph tokens: $\mathcal{T}_{KG}^1 = \mathcal{F}_H(\mathcal{H}_1) \in \mathbb{R}^{P\times d'}$, $\mathcal{T}_{KG}^2 = \mathcal{F}_H(\mathcal{H}_2) \in \mathbb{R}^{P\times d'}$, where d' denotes the dimension of tokens. To represent pairwise discrepancies, we compute a bidirectional residual graph embedding: $\Delta\mathcal{H} = [\mathcal{H}_1 - \mathcal{H}_2; \mathcal{H}_2 - \mathcal{H}_1]$, where $[;]$ denotes the concatenate operation. This residual representation explicitly encodes how the two performers differ in the learned kinematic relation space. We then map the residual graph embedding into discrepancy tokens: $\mathcal{T}_{\Delta KG} = \mathcal{F}_H(\Delta\mathcal{H}) \in \mathbb{R}^{P\times d'}$. These tokens serve as structured evidence for language generation.

Action-difference description generation. To generate the final action-difference description, we concatenate the graph tokens of both videos, the residual discrepancy tokens, and an instruction-prompt embedding f_P: $\hat{\mathcal{C}} = \text{Text_Decoder}([\mathcal{T}_{KG}^1; \mathcal{T}_{KG}^2; \mathcal{T}_{\Delta KG}; f_{\mathcal{P}}])$. The decoder therefore receives both individual kinematic evidence and explicit residual comparison evidence. This design allows the generated description to be grounded in the differences between two performers rather than in isolated video-level representations. The original method similarly generates residual keypoint graph embeddings by contrasting $\mathcal{H}_1$ and $\mathcal{H}_2$, then feeds the corresponding tokens into a text decoder for action-difference description generation. To supervise description generation, we encode both the ground-truth action-difference description C and the generated description $\hat{C}$ using a language model: $f_{\mathcal{C}} = \text{LLM}(\mathcal{C}) \in \mathbb{R}^d$, $f_{\hat{\mathcal{C}}} = \text{LLM}(\hat{\mathcal{C}}) \in \mathbb{R}^d$. We then minimize the semantic discrepancy:

$$\mathcal{L}_{\text{comment}} = 1 - \cos_\text{sim}(f_{\mathcal{C}}, f_{\hat{\mathcal{C}}}). \tag{6}$$

3.5 Learning Objective

The full training objective combines score prediction, semantic grounding, action-difference generation, and cross-video routing consistency:

$$\mathcal{L} = \mathcal{L}_{\text{score}} + \alpha_1\mathcal{L}_{\text{semantic}} + \alpha_2\mathcal{L}_{\text{comment}} + \alpha_3\mathcal{L}_{\text{KL}}, \tag{7}$$

where α_1, α_2 and α_3 are coefficients controlling the weights of loss functions. During inference, expert commentary is not required. Given only two input videos, SKR selects anatomy-grounded kinematic routes, constructs grounded kinematic representations, predicts individual scores, and generates a natural-language description of the performers' action differences.

4 Experiments

4.1 Evaluation Metrics & Datasets

The proposed UADR task contains two sub-tasks: (1) score prediction, also known as action quality assessment (AQA), and (2) action difference description generation (ADDG). We test our model on three datasets: (1) our constructed ADR dataset for UADR; (2) T3SeT dataset Ma et al. (2025) for table tennis performance assessment; (3) FitnessAQA dataset Parmar et al. (2022) for assessing fitness actions in gyms. Detailed formulation of evaluation metrics and introduction of datasets are contained in supplementary materials.

4.2 Quantitative Results

We compare our SKR framework with existing approaches for AQA and foundation models for action score prediction. Results in Table 1 and Table 2 demonstrate that our approach outperforms existing state-of-the-art approach for AQA in action score prediction. In addition, our approach outperforms both open source (LLaVA-Video Zhang et al. (2024), QWen3-VL Bai et al. (2025)) and closed source (GPT-5 OpenAI (2025), Gemini-2.5 Google DeepMind (2025)) large vision-language models (LVLM). For all open source LVLMs, we finetune the model with LoRA Hu et al. (2022)for comparison. For action difference description generation, results in Table 3 demonstrates that our approach is able to generate better descriptions compared with competing approaches.

We also compare action difference description generation results for each category in Table 12 of appendix. For the FitnessAQA dataset, we evaluate four fitness actions: OHP knee, OHP elbow, Squar forward and Squat inward. Results in Table 2 indicate that our SKR approach outperforms

Models	$\rho\uparrow$	$R-\ell_2(\times100)\downarrow$
FineParser Xu et al. (2024c)	0.41	29.36
MAGR Zhou et al. (2024b)	0.73	10.14
Qwen3-VL Bai et al. (2025)	0.46	44.26
LLaVA-Video Zhang et al. (2024)	0.21	16.93
GPT-5 OpenAI (2025)	0.54	11.87
Gemini 2.5 Google DeepMind (2025)	0.61	5.57
Ours (Δ Expert Comment)	0.69	8.78
Our (Δ GNN)	0.74	4.45
Ours (Δ KL loss)	0.67	7.11
Ours (Δ Semantic Loss)	0.63	6.06
Ours (Δ 3D Mesh)	0.65	5.37
Ours	**0.76**	**3.37**

Table 1: Action score prediction results in the ADR dataset. The best results are highlighted in red and second best results are highlighted with underlines.

Models	T3SeT			FitnessAQA			
	Acc	F1	Pre	F1 (OK)	F1 (OE)	F1 (SF)	F1 (SI)
MAGR Zhou et al. (2024b)	0.556	0.262	0.271	0.564	0.357	0.796	0.157
LLaVA-Video Zhang et al. (2024)	0.208	0.115	0.069	0.392	0.533	0.795	0.257
Gemini 2.5 Google DeepMind (2025)	-	-	-	0.105	0.128	0.513	0.170
GPT-5 OpenAI (2025)	-	-	-	0	0.046	0	0
Qwen3-VL Bai et al. (2025)	0.417	0.260	0.304	0.404	0.082	0	0
Ours (Δ KL Loss)	0.736	0.315	0.325	0.598	0.490	0.477	0.213
Ours (Δ 3D mesh)	0.347	0.232	0.361	0.542	0.447	0.485	0.219
Ours	**0.764**	**0.348**	**0.338**	**0.769**	**0.834**	**0.822**	**0.920**

Table 2: Action score prediction results in T3SeT and FitnessAQA datasets. OK: OHP_knee; OE: OHP_elbow; SF: Squat_forward; SI: Squat_inward. Acc: accuracy; F1: F1 score; Pre: precision. Other notations are the same with Table 1.

other approaches by a large margin on the FitnessAQA dataset across all the action categories. For the T3SeT dataset, as there are no ground–truth score provided, we employ QWen3 Bai et al. (2025) organize videos into three different levels (novice, middle, expert) based on the given expert commentary. Results in Table 2 demonstrate that our SKR framework achieve much better performance on the T3SeT dataset across all evaluation metrics. More quantitative rsutls are presented in Section F in the appendix.

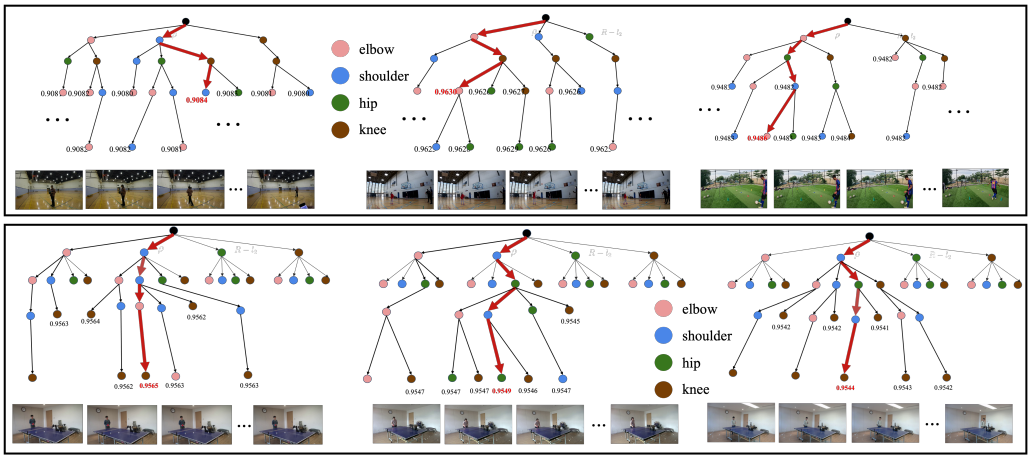

Figure 5: Visualization results of visual reasoning paths in kinematic touting tree. Upper box: ADR dataset for sports actions; Lower box: T3SeT dataset for table tines training. The selected reasoning paths are highlighted in red arrows. Score values of top 10 paths are annotated in the figure.

4.3 Ablation Studies

Ablation studies in both Table 1 and Table 2 demonstrate the effectiveness and necessity of our designed modules and loss functions. Results in Table 1 indicate that removing semantic loss and 3D mesh result in the most significant performance drop in terms of action score prediction. This suggests that language features from expert commentary and fine-grained 3D kinematic cues are essential for guiding visual representations in decision making. More ablation results are provided in Section F in appendix.

Models	BLUE (n_gram)					ROUGE				METEOR
	NG1	NG2	NG3	NG4	All	RO-1	RO-2	RO-L	RO-S	score
LLaVA-Video Zhang et al. (2024)	16.8	1.6	0.2	0.2	0.9	15	11	10.6	10.6	14.4
Gemini 2.5Google DeepMind (2025)	15.8	1.1	0.3	0.2	0.8	18.6	0.9	11.9	11.9	13.1
GPT-5 OpenAI (2025)	21.6	1.8	0.3	0.2	1.1	22.3	1.2	14.6	14.6	16.3
Qwen3-VL Bai et al. (2025)	35.8	11.7	3.5	**1.9**	5.9	37.1	11.4	25.1	25.1	32.1
Ours (Δ Semantic Loss)	31.1	10.2	3.5	1.5	5.6	35.9	11.6	25.1	24.8	28.9
Ours (Δ GNN)	32.8	10.3	3.7	1.6	5.9	35.3	11.0	24.4	24.4	29.9
Ours (Δ RKG)	32.8	9.9	3.0	1.2	3.7	36.0	11.0	24.1	24.1	30.1
Ours (Δ KL Loss)	35.2	10.9	3.6	1.5	5.9	35.9	10.9	23.5	23.5	32.4
Ours (Δ 3D Mesh)	34.5	10.6	3.7	1.5	6.0	36.6	11.0	25.0	25.0	31.7
Ours	**37.0**	**12.1**	**3.8**	1.6	**6.2**	**37.8**	**12.0**	**25.3**	**25.3**	**33.0**

Table 3: Quantitative results for action difference description generation on the ADR dataset. Δ RKG: residual keypoint graph. NG: n_gram; RO: ROUGE. Other notations are the same with Table 1.

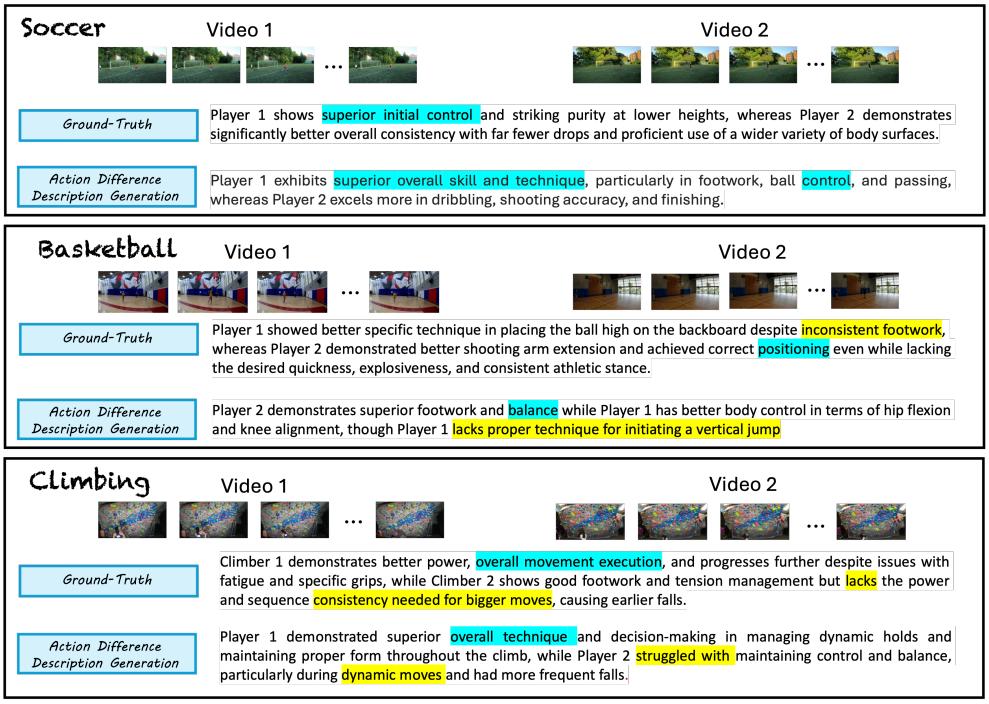

Figure 6: Qualitative results of generated action difference descriptions in our ADR dataset. Three sports categories (Soccer, Basketball, Climbing) with two videos forming a video pair. Highlighted texts indicates the similar semantic descriptions from our generated text with ground-truth.

4.4 Qualitative Results

Figure 5 visualizes the learned reasoning paths from kinematics-aware trees. For those paths contain nodes belong to the same keypoint, we combine them into one single node. This visualization effectively demonstrates how our SKR approach successfully models expert visual reasoning processes, capturing the hierarchical and sequential nature of domain-specific assessment protocols. Figure 6 presents qualitative comparisons between generated action difference descriptions and ground-truth annotations. The highlighted text segments demonstrate that our generation framework effectively captures critical assessment information present in the ground-truth, validating the semantic fidelity of our approach in articulating visual differences that domain experts would identify as significant. More qualitative results are presented in appendix.

5 Conclusion

In this paper, we introduced Unified Action Difference Reasoning (UADR), a task that jointly predicts performance scores and explains fine-grained differences between paired action videos. To address UADR, we proposed Structured Kinematic Routing (SKR), which models comparative action assessment through anatomy-grounded kinematic routes, cross-modal semantic grounding, and residual kinematic discrepancy reasoning. We also constructed ADR, a paired-video benchmark with expert-derived scores and action-difference descriptions across multiple skilled activities. Experiments show that SKR improves both score prediction and action-difference generation over prior action quality assessment methods and large vision-language baselines. These results demonstrate that structured reasoning over human kinematics provides an effective inductive bias for fine-grained, interpretable action comparison. The primary limitation of this work is its focus on the sports domain, which represents a relatively narrow scope. Future work will extend SKR to broader domains such as rehabilitation, surgical training, and education, where comparative feedback and actionable explanations are critical.

AI DISCLOSURE

In this work, we used generative AI tools to (1) aid or polish writing, (2) find related work, (3) generate synthetic datasets. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kumar Ashutosh and Kristen Grauman. Learning skill-attributes for transferable assessment in video. *arXiv preprint arXiv:2511.13993*, 2025.

Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. Expertaf: Expert actionable feedback from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2408.00672.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European conference on computer vision*, pp. 422–438. Springer, 2022.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

James Burgess, Xiaohan Wang, Yuhui Zhang, Anita Rau, Alejandro Lozano, Lisa Dunlap, Trevor Darrell, and Serena Yeung-Levy. Video action differencing. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2503.07860.

Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. URL https://arxiv.org/abs/2406.06087.

Google DeepMind. Gemini 2.5. *https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking*, 2025.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Mingzhe Li, Hong-Bo Zhang, Qing Lei, Zongwen Fan, Jinghua Liu, and Ji-Xiang Du. Pairwise contrastive learning network for action quality assessment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. doi: 10.1007/978-3-031-19772-7_27.

Wenhao Li, Mengyuan Liu, Hong Liu, Pichao Wang, Jialun Cai, and Nicu Sebe. Hourglass tokenizer for efficient transformer-based 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 604–613, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Ji Ma, Jiale Wu, Haoyu Wang, Yanze Zhang, Xiao Xie, Zheng Zhou, Hui Zhang, Jiachen Wang, and Yingcai Wu. T3set: A multimodal dataset with targeted suggestions for llm-based virtual coach in table tennis training. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5686–5697, 2025.

Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. RICA^2: Rubric-informed, calibrated assessment of actions. In *European Conference on Computer Vision*. Springer, 2024. URL https://arxiv.org/abs/2408.02138.

Filip Noworolnik and Joanna Jaworek-Korjakowska. Assessing the quality of soccer shots from single-camera video with vision-language models and motion features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2754–2761, 2025. doi: 10.1109/ICCVW69036.2025.00287.

OpenAI. Gpt-5. *https://openai.com/gpt-5/*, 2025.

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*, pp. 6331–6340, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 304–313, 2019.

Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 105–123. Springer, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/2006.07665.

Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, pp. 4902–4910, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3474085.3475438. URL https://arxiv.org/abs/2201.03746.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.

Chi Hsuan Wu, Kumar Ashutosh, and Kristen Grauman. Skillsight: Efficient first-person skill assessment with gaze. *arXiv preprint arXiv:2511.19629*, 2025.

Huangbiao Xu, Xiao Ke, Yuezhou Li, Rui Xu, Huanqi Wu, Xiaofeng Lin, and Wenzhong Guo. Vision-language action knowledge learning for semantic-aware action quality assessment. In *European Conference on Computer Vision*, pp. 423–440. Springer, 2024a.

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 (CVPR)*, pp. 2949–2958, 2022a. URL https://arxiv.org/abs/2204.03646.

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*, pp. 2949–2958, 2022b.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14628–14637, 2024b. URL https://arxiv.org/abs/2405.06887.

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*, pp. 14628–14637, 2024c.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2108.07797.

Wulian Yun, Mengshi Qi, Fei Peng, and Huadong Ma. Semi-supervised teacher-reference-student architecture for action quality assessment. In *European Conference on Computer Vision*. Springer, 2024. URL https://arxiv.org/abs/2407.19675.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Kanglei Zhou, Liyuan Wang, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Jianguo Li, and Xiaohui Liang. Magr: Manifold-aligned graph regularization for continual action quality assessment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a. URL https://arxiv.org/abs/2403.04398.

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*. Springer, 2024b. URL https://arxiv.org/abs/2403.04398.

A Related Work

Qualitative Action Quality Assessment. Recent advancements in qualitative action quality assessment generate actionable feedback that explicitly guides user improvement. ExpertAF Ashutosh et al. (2025), SkillSight Wu et al. (2025) and SportSkills Ashutosh & Grauman (2025) pioneered this direction by introducing multimodal frameworks that outputs both free-form textual commentary and synthesized visual demonstrations of corrected motions. Building on this, CrossTrainer Ashutosh & Grauman (2025) addresses the challenge of generalization by learning "skill-attributes", which are semantic concepts such as balance, control, and stance, that transcend specific sports boundaries. Video action differencing Burgess et al. (2025) proposes an argentic workflow to qualitatively identify differences between videos performing the same action.

Quantitative Action Quality Assessment. Quantitative action quality assessment has gained considerable attention recently, with diverse approaches aimed at enhancing accuracy and robustness. Wang et al. Tang et al. (2020) propose an uncertainty-aware score distribution learning framework, while Yang et al. Yu et al. (2021) introduce a group-aware contrastive regression model leveraging group dynamics for refined assessments. The TSA-Net architecture by Wang et al. Wang et al. (2021) uses tube self-attention to focus on relevant action segments, achieving improved performance. Additionally, FineDiving Xu et al. (2022a) provides a fine-grained, procedure-aware dataset for training models on nuanced action quality data. Further approaches include Wang et al.'s pairwise contrastive learning network Li et al. (2022), which emphasizes relationships between action pairs, and the FineParser model Xu et al. (2024b), which enables fine-grained spatio-temporal parsing for human-centric evaluations. The MAGR method Zhou et al. (2024a) uses manifold-aligned graph regularization for continual assessment, adapting to evolving data distributions, while the GAIA framework Chen et al. (2024) tackles the unique challenges of AI-generated video assessment. These studies collectively highlight the importance of robust frameworks and datasets in advancing action quality assessment. Recent work by Burgess et al. Burgess et al. (2025) introduced the video action differencing task, which compares how two individuals perform the same actions. However, their approach is limited to qualitative assessment only. Recent work on assessing soccer shots Noworolnik & Jaworek-Korjakowska (2025) is able to provide expert-style feed- back and scalar ratings, but is constrained to one sport category and lacks of modeling structured expert attention.

In contrast to existing work, we addresses both *quantitative* and *qualitative* assessment by modeling expert cognitive process, enabling more comprehensive evaluation of performer differences.

B Evaluation Metrics

For AQA evaluation, we employ two standard metrics: Spearman's rank correlation (ρ) and relative ℓ_2 distance R-ℓ_2. The ADDG performance is assessed using established natural language generation metrics: BLEU Papineni et al. (2002), ROUGE Lin (2004) and METEOR Banerjee & Lavie (2005), reported out of 100, higher is better. Among these metrics, R-ℓ_2 is unique in that lower values indicate better performance, while higher values of ρ, BLEU, METEOR and ROUGE signify superior performance. We test our model on three datasets: (1) our constructed ADR dataset for sports action assessment; (2) T3SeT dataset Ma et al. (2025) for table tennis performance assessment; (3) FitnessAQA dataset Parmar et al. (2022) for assessing fitness actions in gyms. The assessment of models on the FitnessAQA dataset is evaluated with F1 score following the practice of previous work Parmar et al. (2022). Evaluation on the T3SeT dataset employ precision, accuracy and F1 score. Detailed formulation of evaluation metrics are presented as follows:

(I) For score prediction on the ADR dataset, we utilize Spearman's rank correlation (ρ) and relative ℓ_2 distance $R - \ell_2$ Pan et al. (2019); Parmar & Morris (2019) as evaluation metrics, while on the FitnessAQA dataset, we use F1-score as the evaluation metric:

- **Spearman's rank correlation (ρ)** focuses on testing sample ranking and is defined as :

$$\rho = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{8}$$

  where y and $\hat{y}$ indicate the ranking of two sequences. Higher ρ suggests better model performance.

- **Relative ℓ_2 distance** ($R-\ell_2$) measures the distance from each sample to the ground-truth:

$$R{-}\ell_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|y_i - \hat{y}_i|}{y_{\text{max}} - y_{\text{min}}}\right), \tag{9}$$

  where y_i and $\hat{y}_i$ denote the ground-truth and prediction scores for the i-th sample, respectively. Lower R-ℓ_2 indicates better performance.

- **F1-score** (F1) is the harmonic mean of precision and recall:

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{10}$$

  where

(II) For action difference description generation, we employ BLEU Papineni et al. (2002), ROUGE Lin (2004) and METEOR Banerjee & Lavie (2005) as evaluation metrics:

- **BLEU (Bilingual Evaluation Understudy)** measures how many words or phrases from a generated (candidate) translation overlap with a set of reference translations. It focuses on precision and considers n-grams to evaluate fluency and adequacy.
- **METEOR (Metric for Evaluation of Translation with Explicit Ordering)** is designed to address some of BLEU's limitations by considering recall alongside precision and focusing on word-level alignment. It also incorporates semantic similarity through synonyms and stemming.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** is a family of metrics primarily used to evaluate automatic text summarization systems, though it also applies to machine translation. ROUGE compares automatically generated summaries or translations against one or more human-produced reference texts. All ROUGE metrics produce scores between 0 and 1, where higher values indicate greater similarity between the system output and human references.

C More on the ADR Dataset Construction

ADR Dataset Statistics. Detailed statistics on each sports category of the ADR dataset are presented in Table 4.

Category	Sub-activity	# Videos		# Video Pairs	
		train	test	train	test
Basketball	Mid-Range Shooting	17	28	136	378
	Mikan Layup	26	23	325	253
	Reverse Layup	27	30	351	435
Soccer	Juggling	6	2	15	1
	Penalty Kick	21	3	210	2
	Dribbling	23	5	253	10
Rock Climbing	Climbing	70	155	2415	703

Table 4: Dataset Splits for each sports category. Both basketball and soccer contain three categories, while rock climbing contains only one category. The number of videos and video pairs for each sports category are listed.

Human evaluation. To validate the quality of action difference descriptions generated by Gemini 2.5, we conducted a systematic human evaluation. We select 300 randomly sampled video pairs from each sport category. Three human evaluator with domain knowledge in each respective sport was tasked with determining which athlete demonstrated superior performance, based solely on the model-generated action difference descriptions. We then compared these human judgments against the ground-truth athlete scores to assess whether the generated descriptions accurately captured performance differentials. The agreement rates are as follows: H1 (868/900, 96.4%), H2 (857/900,

95.2%), and H3 (852/900, 94.7%). These consistently high agreement scores indicate that the generated action difference descriptions are reliable and accurately capture fine-grained performance variations between athletes.

Furthermore, we validate whether the described attributes (who does which part better) are correct. The described attributes include visual correctness, performer attribution, source support, and unsupported inference. Specifically, we sample 300 video pairs, splitted equally across the three sports and balanced by score difference and source-comment coverage. We additionally evaluate a targeted subset containing asymmetric comments, where an attribute is mentioned for only one performer. Each description is decomposed into atomic comparative claims, such as "`performer 1 maintains better knee alignment than performer 2`". One annotator performs this decomposition and three domain-informed raters then independently verify it by evaluating every claim in two separate passes:

- **Source-support evaluation:** Raters see the claim and both sets of expert comments, but not the scores. Each claim is labeled as: bilaterally supported; supported for only one performer; unsupported for both performers; conflicting; unclear. An attribute omitted from one performer's comments is treated as unknown, not as evidence of inferior performance.
- **Visual-faithfulness evaluation:** Raters see the claim and both videos with randomized performer labels, without seeing the source comments or scores. They determine whether the claimed difference is visually supported, contradicted, or not reliably observable because of viewpoint, occlusion, or ambiguity.

A claim is considered validated only when it has bilateral source support or is independently verified from both videos by at least two of the three raters. Claims based solely on asymmetric omission does not pass this criterion. Ties and unclear cases are reviewed by an additional domain expert. We report bilateral-support rate, visually verified rate, contradiction rate and asymmetric-omission rate, both overall and by sport. The high Bilateral-support rate (89%) and Visually verified rate (91%) in Table 5, together with the low Contradiction rate (2%) and Attribution-reversal rate (2%) demonstrate the reliability generated action difference descriptions.

Activity	Bilateral Support	Visual Verification	Contradiction	Asym. Omission
Soccer	87%	91%	3%	1%
Basketball	92%	88%	1%	1%
Climb	88%	93%	2%	3%
Average	**89%**	**91%**	**2%**	**2%**

Table 5: Consistency statistics across activity categories.

Dataset design choice. To design the UADR task, there is an intuitive approach, which is to compare each performer against a prototypical expert execution (pairwise supervision versus an expert prototype). However, although expert-prototype comparison is useful, it is not a substitute for the task studied here. UADR is intentionally defined on two arbitrary videos of the same action and requires two individual scores plus a description of relative strengths and weaknesses. Fixing one input to a canonical expert would answer a narrower question and **would not support direct comparison between two non-expert executions or feedback on strengths on both sides**. The pairwise setting is therefore a deliberate task choice.

D T3SeT and FitnessAQA Datasets

T3SeT Dataset. The T3SeT dataset Ma et al. (2025) captures seven core table tennis techniques: attack, block, flick, pendulum, push, short, and topspin. In total, T3Set contains data from 32 amateur players across 380 multi-ball training rounds, comprising 8,655 video clips (individual strokes) and 8,395 items of professional coaching advice.

FitnessAQA Dataset. FitnessAQA Parmar et al. (2022) includes three exercises over 5,000 videos covering three gym exercises: 1) BackSquat; 2) BarbellRow; and 3) Overhead (shoulder) Press. It also provides annotations from experts to evaluate the performance of athletes.

E Implementation Details

Feature dimension $d = 4096$. $\mathcal{F}_M, \mathcal{F}_V, \mathcal{F}_K, \mathcal{F}_H, \mathcal{F}_{ME}$ are all implemented with two-layer MLP. Multi-head attention consists of 8 heads. All experiments are conducted on two NIDIA L40S GPUs. Human 3D mesh is extracted from SAM 3D Body Yang et al. (2024). 3D keypoints are extracted from the HOT Li et al. (2024) model. Each input video is divided into $T = 64$ temporal segments of equal duration. From each segment, we sample a frame and resize it to 448×448.Visual features are extracted using the video encoder of InternVideo 2.5 (8B) Wang et al. (2024). To capture 3D structural information, we employ the HOT method to obtain 3D keypoints. To generate text embedding, we use T5 Raffel et al. (2020) as text encoder, due to its lightweight architecture and computational efficiency, enabling fast and scalable extraction of high-quality textual embeddings without introducing significant overhead. We use Qwen3 (7B) Bai et al. (2025) as the text decoder to summarize the differences between two players, as its stronger generative capacity yields more coherent, detailed, and context-aware descriptions compared to smaller language models, which is critical for fine-grained comparative reasoning. Model training is performed using the SGD optimizer with a learning rate of 0.0003.

F More Quantitative Results

Robustness to noisy or imperfect motion signals. To evaluate the model's robustness to noise or missing keypoints/meshes, we *inject noise into 3D keypoints/meshes* as following:

$$f_{\text{keypoint}} \leftarrow f_{\text{keypoint}} + \sigma \cdot \text{std}(f_{\text{keypoint}}) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$
$$f_{\text{mesh}} \leftarrow f_{\text{mesh}} + \sigma \cdot \text{std}(f_{\text{mesh}}) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

where noise ϵ is sampled from a standard gaussian distribution, σ controls the noise intensity, f_{keypoint} and f_{mesh} represent extracted 3D keypoint and mesh. Table 6 presents different noise injection strategies during inference: adding noise to (1) 3D keypoint (KP), (2) 3D mesh, (3) 3D keypoint + 3D mesh. Results in Table 6 demonstrate that adding noise to 3D keypoints or meshes will not undermine model performance too much, indicating that our model performance is robust against noisy 3D keypoints and meshes.

3D KP	$\rho\uparrow$	$R-l_2(\times100)\downarrow$	**3D Mesh**	$\rho\uparrow$	$R-l_2(\times100)\downarrow$	**3D KP&Mesh**	$\rho\uparrow$	$R-l_2(\times100)\downarrow$
Ours ($\sigma = 0.00$)	**0.765**	**3.37**	**Ours ($\sigma = 0.00$)**	**0.765**	**3.37**	**Ours ($\sigma = 0.00$)**	**0.765**	**3.37**
Ours ($\sigma = 0.05$)	0.764	3.39	Ours ($\sigma = 0.05$)	0.764	3.40	Ours ($\sigma = 0.05$)	0.762	3.41
Ours ($\sigma = 0.10$)	0.763	3.40	Ours ($\sigma = 0.10$)	0.760	3.47	Ours ($\sigma = 0.10$)	0.758	3.48
Ours ($\sigma = 0.20$)	0.754	3.57	Ours ($\sigma = 0.20$)	0.753	3.60	Ours ($\sigma = 0.20$)	0.749	3.68
Ours ($\sigma = 0.50$)	0.732	4.12	Ours ($\sigma = 0.50$)	0.740	3.92	Ours ($\sigma = 0.50$)	0.727	4.39

Table 6: Impact of noise injection strategies on 3D keypoints (KP), 3D Mesh, and 3D KP&Mesh.

More ablation studies. We report additional ablation studies in Table 7 that combine multiple key components in the SKR framework. The substantial degradation in ρ and increase in $R - l_2$ across both ablations underscores the effectiveness and necessity of our proposed components. Specifically, removing both the GNN and 3D mesh modules leads to a 0.42 drop in ρ (0.74 $\rightarrow$ 0.32), which is markedly larger than removing the GNN alone. This result indicates that 3D mesh representations provide critical structural and kinematics-aware cues that cannot be compensated for by graph reasoning alone, and are essential for accurate and robust score prediction. Similarly, jointly removing KL loss and Semantic loss results in a 0.31 decrease in ρ (0.67 $\rightarrow$ 0.35), compared to removing KL loss alone. This highlights the importance of the semantic loss in cross-modal semantic grounding, enabling the model to capture semantically meaningful discrepancies between players. Without this alignment, the model's ability to perform fine-grained comparative reasoning is significantly impaired, leading to notable performance degradation.

Collectively, these additional ablation study results demonstrate that both the structural modeling components and the cross-modal alignment objectives are crucial for achieving strong performance.

Runtime overhead. We report the runtime overhead of major components in our framework in the table below, evaluated over 100 videos in total. Runtime overhead of Kinematic Routing Tree and Model Forward Process are included. Results in Table 8 demonstrate that our approach is computationally efficient, with an overall runtime of less than 45 ms.

	$\rho \uparrow$	$R - l_2(\times \mathbf{100}) \downarrow$
Ours (Δ GNN + Δ 3D Mesh)	0.32	7.34
Ours (Δ KL loss + Δ Semantic loss)	0.354	7.46
Ours	**0.765**	**3.37**

Table 7: More ablation results that remove multiple key components.

	Test Time
Kinematic Routing Tree	8.51 ± 0.02 ms
Model Forward Process	36.42 ± 0.14 ms
Total	$\mathbf{44.93 \pm 0.14}$ **ms**

Table 8: Runtime overhead of the SKR framework in the ADR dataset. Results are presented in the form of mean $\pm$ std over 100 videos.

Impact of different kinematic routing tree depth. In Table 9 we present the ablation studies of implementing kinematic routing tree with different depth. The results indicate that 4-layer kinematic routing tree result in best score prediction performance.

	$\rho \uparrow$	$R - \ell_2\ (\times\ 100) \downarrow$
Ours (3 layers)	0.70	4.56
Ours (5 layers)	0.36	10.41
Ours (4 layers)	**0.76**	**3.37**

Table 9: Ablation studies on the effect of different depth of tree search on score prediction performance.

Score prediction performance on the ADR dataset for each category. In Table 11 and Table 12, we present detailed results of score prediction performance on individual sports category (ADR dataset). We show detailed quantitative results of action difference description generation of individual sports category in Table 12.

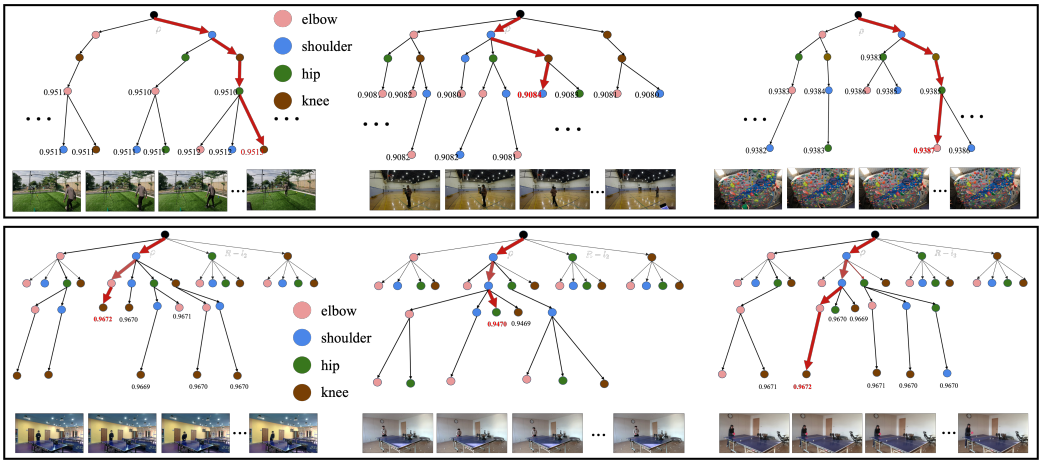

Figure 7: More qualitative visualization results of visual reasoning paths in kinematic touting tree. Upper box: ADR dataset for sports actions; Lower box: T3SeT dataset for table teens training. The selected reasoning paths are highlighted in red arrows. score values of top 10 paths are annotated in the figure.

Effect of the implicit-graph threshold. The threshold applies only to the learned graph over non-adjacent keypoint pairs. It does not remove the physical skeleton graph, which remains a separate input. Lowering the threshold makes the auxiliary graph denser, retaining weaker long-range

Models	BLUE					ROUGE				METEOR	$\rho\uparrow$	$R-\ell_2(\times100)\downarrow$
	NG1	NG2	NG3	NG4	All	RO-1	RO-2	RO-L	RO-S	score		
Soccer	0.399	0.116	0.032	0.013	0.058	0.391	0.125	0.268	0.268	0.334	0.93	1.51
Basketball	0.375	0.119	0.035	0.015	0.061	0.381	0.119	0.275	0.275	0.319	0.85	3.37
Climbing	0.367	0.120	0.040	0.017	0.064	0.373	0.120	0.257	0.257	0.321	0.74	5.21
All	**0.369**	**0.119**	**0.038**	**0.016**	**0.063**	**0.384**	**0.124**	**0.269**	**0.269**	**0.332**	**0.80**	**3.55**

Table 10: Quantitative results of our approach for action difference description generation and action score prediction on individual sports categories in the ADR dataset. NG: n_gram; RO: ROUGE.

Models	Soccer		Basketball		Climbing	
	$\rho\uparrow$	$R-\ell_2\ (\times100)\downarrow$	$\rho\uparrow$	$R-\ell_2\ (\times100)\downarrow$	$\rho\uparrow$	$R-\ell_2\ (\times100)\downarrow$
MAGR Zhou et al. (2024b)	0.52	8.43	0.72	7.90	0.74	13.32
FineParser Xu et al. (2024c)	0.71	45.52	0.12	24.43	0.72	13.18
LLaVA-Video Zhang et al. (2024)	0.07	29.64	0.31	13.58	0.48	20.23
GPT-5 OpenAI (2025)	0.18	38.64	0.51	10.17	0.75	16.7
Gemini 2.5 Google DeepMind (2025)	0.64	14.21	0.63	7.01	0.70	4.34
Qwen 3 Yang et al. (2024)	0.70	69.49	0.48	11.37	0.43	59.67
Ours (Δ Expert Commentary)	0.57	8.72	0.77	4.68	0.56	6.37
Ours (Δ GNN)	0.48	7.26	0.75	5.99	0.54	7.14
Ours (Δ KL loss)	0.75	4.15	0.80	4.52	0.70	4.90
Ours (Δ Semantic Loss)	0.65	9.612	0.78	9.18	0.51	5.59
Ours	**0.93**	**1.51**	**0.85**	**3.37**	**0.74**	**5.21**

Table 11: Quantitative results of score prediction performance on the ADR dataset for the three sports categories (Soccer, Basketall, Climbing).

Models	BLUE					ROUGE				METEOR
	NG=1	NG=2	NG=3	NG=4	all	RO-1	RO-2	RO-L	RO-S	score
LLaVA-Video (Soccer) Zhang et al. (2024)	15.5	1.5	0.2	0.2	0.7	14.3	0.6	9.3	9.3	12.6
LLaVA-Video (Basketball) Zhang et al. (2024)	16.3	1.7	0.2	0.2	0.9	13.6	0.5	9.9	9.9	14.5
LLaVA-Video (Climbing) Zhang et al. (2024)	16.9	1.6	0.2	0.2	0.9	15.4	1.2	1.1	1.1	14.3
GPT-5 (Soccer) OpenAI (2025)	23.8	2.1	0.6	0.6	1.3	23.5	1.6	16.0	16.0	17.1
GPT-5 (Basketball) OpenAI (2025)	22.0	1.7	0.3	0.3	1.1	22.5	1.4	14.7	14.7	15.8
GPT-5 (Climbing) OpenAI (2025)	21.2	1.8	0.3	0.2	1.1	22.1	1.2	14.5	14.5	16.5
Gemini 2.5 (Soccer) Google DeepMind (2025)	21.7	2.2	0.2	0.2	1.1	26.7	2.6	14.8	14.8	17.0
Gemini 2.5 (Basketball) Google DeepMind (2025)	16.3	1.0	0.3	0.2	0.8	18.0	1.0	11.7	11.7	13.0
Gemini 2.5 (Climbing) Google DeepMind (2025)	15.5	1.1	0.3	0.2	0.8	18.6	0.9	12.0	12.0	13.0
QWen3-VL (Soccer) Bai et al. (2025)	35.9	15.4	3.2	2.7	5.4	39.4	13.9	29.1	29.1	33.9
QWen3-VL (Basketball) Bai et al. (2025)	35.8	13.4	4.9	1.9	5.9	36.1	10.9	24.4	24.4	32.6
QWen3-VL (Climbing) Bai et al. (2025)	35.7	10.2	3.5	1.8	5.4	37.1	12.7	26.8	26.8	33.3
Ours (Δ Expert Commentary) (Soccer)	36.4	9.6	2.4	0.8	4.5	37.1	10.2	26.1	26.1	31.1
Ours (Δ Expert Commentary) (Basketball)	36.6	10.7	3.4	1.4	5.8	37.7	11.4	26.6	26.6	30.6
Ours (Δ Expert Commentary) (Climbing)	34.2	10.9	3.9	1.8	6.3	36.8	11.6	25.3	25.3	32.1

Table 12: Quantitative results of different approaches for action difference description generation performance over individual sports categories on the ADR dataset. NG: n_gram; RO: ROUGE.

dependencies but admitting more noise and possible oversmoothing. Raising it makes the graph sparser, improving selectivity but potentially removing useful coordination cues. Eq 4 specifies the learned probabilities and thresholding mechanism. Threshold in Eq 4 is 0.5. We include two additional threshold (τ) value 0.3 and 0.7 and list experiment results in Table 13. Results in Table 13 demonstrate that threshold 0.5 would obtain the best performance in most cases.

G More Qualitative Results

We show more visualization results of the reasoning path in keypoint-guided tree in both the ADR dataset and T3SeT dataset in Figure 7. In Figure 8, we present more qualitative results of generated action difference descriptions in the ADR dataset.

ADR Dataset	$\rho \uparrow$	$R - l_2\,(\times 100) \downarrow$
$\tau = 0.3$	0.25	7.90
$\tau = 0.7$	0.19	7.93
$\tau = 0.5$ **(Ours)**	**0.76**	**3.37**

FitnessAQA Dataset	F1 Score (OK)	F1 Score (OE)	F1 Score (SF)	F1 Score (SI)
$\tau = 0.3$	0.769	**0.845**	0.812	0.918
$\tau = 0.7$	0.769	0.844	0.800	0.920
$\tau = 0.5$ **(Ours)**	**0.769**	0.834	**0.822**	**0.920**

Table 13: Ablation study of different threshold values τ on the ADR and FitnessAQA datasets. Best results are highlighted in red.

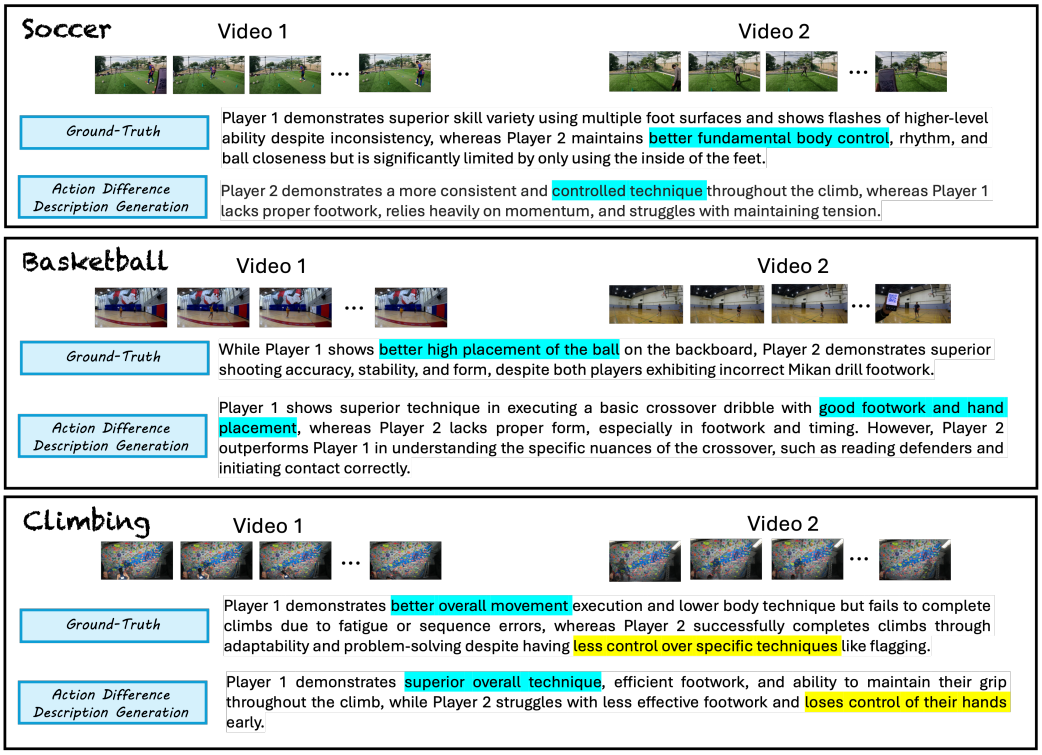

Figure 8: More qualitative results of generated action difference descriptions in our ADR dataset. Three sports categories (Soccer, Basketball, Climbing) with two videos forming a video pair. Highlighted texts indicates the similar semantic descriptions from our generated text with ground-truth.