LEARNING FROM WHAT IT LEARNED: DECODING POST-CONVERGENCE STATES FOR UNIFIED ACTION QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified action quality assessment (AQA) trains a single model across heterogeneous sports, enabling cross-sport representation sharing while converging to a shared solution shaped by heterogeneous sport-specific preferences. Post-convergence analysis reveals limited directional agreement among sport-specific gradients and substantial variation in dominant local curvature around the shared solution, while the raw-to-EMA trajectory contains an empirically favorable interior weight state. Motivated by these observations, we propose **TRAILS**, a post-convergence decoding framework that exploits the model states produced during unified training in weight and statistics spaces. In weight space, **Trajectory Interpolation (TI)** searches the raw-to-EMA path for a favorable shared weight configuration. In statistics space, **Statistics Recalibration (SR)** re-estimates sport-specific BatchNorm statistics from unlabeled test samples, with sport identities inferred automatically from frozen visual features. TRAILS retains a single shared AQA weight configuration without gradient-based retraining of the AQA model. Experiments on three heterogeneous AQA benchmarks show that TRAILS achieves an average SRCC of 0.777 and an R-ℓ_2 of 2.008, improving the overall performance of unified AQA.

1 INTRODUCTION

Action quality assessment (AQA) (Zhou et al., 2024b; Ji et al., 2023; Zhang et al., 2023; Xu et al., 2024b; Li et al., 2024) aims to quantify human performance from video, supporting applications in sports analytics (Xu et al., 2024a), medical rehabilitation (Liu et al., 2021b), and skill determination (Yu et al., 2026). Recent methods (Xu et al., 2022a; Zhou et al., 2023; Ke et al., 2024; Zeng & Zheng, 2024; Zhou et al., 2024a; Liu et al., 2025; Zhou et al., 2025) have achieved strong performance on individual benchmarks. However, AQA across heterogeneous sports, such as rhythmic gymnastics, diving, and lacrosse, remains challenging because these sports differ substantially in motion patterns, temporal structure, and scoring criteria. Training an independent model for each sport limits cross-sport parameter sharing and increases deployment cost. Unified AQA models (Xu et al., 2026c; 2025b;a) instead learn a single model across multiple sports, enabling cross-sport representation sharing within a unified inference framework.

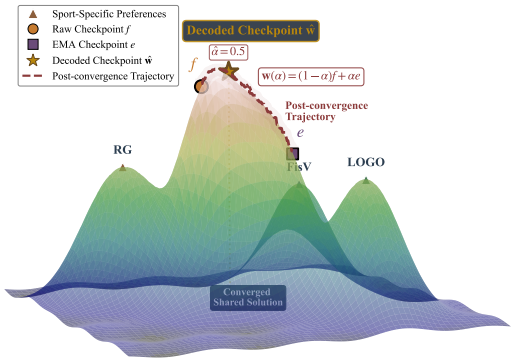


Figure 1: Toy illustration of the raw-to-EMA interpolation space in unified AQA. The raw checkpoint f and EMA companion e define an interpolation path containing accessible interior states.

A shared model must nevertheless accommodate heterogeneous sport-specific preferences during optimization. Different sports may favor different parameter configurations, and their optimization

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.

signals may exhibit limited directional agreement even at convergence, as studied in multi-task optimization (Yu et al., 2020; Liu et al., 2021a; Chen et al., 2018). Our post-convergence analysis characterizes the dominant first- and second-order optimization geometry around the shared solution. Sport-specific gradients remain nonzero and exhibit limited directional agreement, while the dominant Hessian curvature varies substantially across sports (Table 5). These findings characterize the heterogeneous optimization signals present near the shared checkpoint and motivate us to examine the model states retained at the end of unified training.

In particular, we revisit the final raw checkpoint $\mathbf{f}$ and its exponential moving average (EMA) companion $\mathbf{e}$, which summarize the terminal training trajectory differently, as illustrated in Fig. 1. Prior studies of mode connectivity have shown that neural network solutions can be connected through low-loss regions in parameter space (Garipov et al., 2018; Draxler et al., 2018; Frankle et al., 2020). We examine the linear interpolation between the raw checkpoint and its EMA companion and observe that the training-side score objective remains low along the evaluated path, reaching its lowest evaluated value at an interior state. Local geometric measurements further characterize the raw-to-EMA displacement through its alignment with sport-specific gradients and dominant Hessian eigendirections. Importantly, the value of the interior state is established through direct evaluation of the interpolation path and its predictive performance, rather than inferred from local alignment measurements. These findings motivate us to treat the raw-to-EMA segment as a post-convergence decoding space.

Based on this observation, we propose **TRAILS**, a post-convergence decoding framework that operates in weight and statistics spaces. In weight space, **Trajectory Interpolation (TI)** performs a one-dimensional search along the raw-to-EMA path, selecting a shared weight configuration using the training-side score objective. In statistics space, **Statistics Recalibration (SR)** re-estimates sport-specific BatchNorm statistics for the decoded weights from unlabeled test samples, with sport identities inferred automatically from frozen visual features. SR thus provides sport-specific normalization states while retaining the same shared AQA weights. Together, TI and SR decode a post-convergence model state without gradient-based retraining of the AQA model or modification of its underlying architecture.

Experiments on three heterogeneous AQA benchmarks show that TRAILS achieves an average SRCC of 0.777 and an R-ℓ_2 of 2.008, improving overall performance relative to the evaluated unified AQA baseline (Xu et al., 2026a). Additional analyses examine the distribution of the raw-to-EMA displacement across model components and localize the performance gain from TI. Qualitative visualizations further illustrate changes in learned representations and provide sample-level examples of the resulting score predictions.

In summary, we make the following contributions.

- We characterize limited directional agreement among sport-specific gradients and variation in dominant curvature, and identify a favorable interior state along the raw-to-EMA path.
- We propose **TRAILS**, a post-convergence decoding framework that combines Trajectory Interpolation and Statistics Recalibration to select a shared weight configuration and re-estimate sport-specific normalization states without retraining the AQA model.
- Experiments on three heterogeneous AQA benchmarks demonstrate improvements over the evaluated unified AQA baselines. Trajectory analysis, component-level interpolation, and representation visualization further characterize the post-convergence decoding process.

2 Related Work

Action Quality Assessment. Action quality assessment (AQA) estimates action execution quality from video. Early learning-based methods primarily relied on spatio-temporal representation learning followed by continuous score regression (Parmar & Morris, 2019b; Tang et al., 2020; Pan et al., 2019). Subsequent approaches introduced temporal structure modeling and relative assessment to better capture quality variation across performances. Temporal parsing and hierarchical graph models characterize long-range dependencies and structured motion semantics (Bai et al., 2022; Zhou et al., 2023), while pairwise and group-aware contrastive regression explicitly model inter-sample quality relations (Li et al., 2022; Yu et al., 2021; Ke et al., 2024).

For long-term and structurally complex actions, recent work has increasingly focused on finer temporal organization and scoring structure. GDLT decomposes continuous scores into discrete quality grades (Xu et al., 2022a), while FineDiving introduces procedure-aware assessment through alignment of fine-grained action steps (Xu et al., 2022b). Related studies further investigate fine-grained spatio-temporal parsing (Gedamu et al., 2023; Xu et al., 2024b), coarse-to-fine quality modeling (Zhou et al., 2024a), and interpretable or rule-aware assessment (Okamoto & Parmar, 2024; Li & Leung, 2026). Multimodal methods additionally incorporate audio, pose, and language cues to enrich quality representations (Zeng & Zheng, 2024; Xia et al., 2023; Xu et al., 2025a; 2026b). Despite these advances, most existing methods are developed under sport-specific or benchmark-specific settings with comparatively homogeneous motion patterns and scoring semantics.

Unified Action Quality Assessment. Unified AQA extends sport-specific assessment to heterogeneous sports within a single shared model. Early studies considered cross-action assessment and knowledge transfer across multiple action categories (Parmar & Morris, 2019a), while recent methods emphasize shared representation learning and adaptive specialization for heterogeneous assessment (Xu et al., 2026a; 2025b;a). MoAKE introduces a mixture of action knowledge experts to model diverse action patterns within a shared semantic space (Xu et al., 2026a). Segment-aware prototypes accommodate different temporal structures, while adaptive intra- and inter-segment modeling and soft routing provide input-dependent specialization. Related multimodal approaches further incorporate vision-language and audio-visual information for semantic modeling across action domains (Xu et al., 2025a; Zeng & Zheng, 2024).

Existing unified AQA methods primarily address cross-sport heterogeneity through architectural specialization and adaptive representation learning. TRAILS instead operates on the states produced by unified optimization after convergence and decodes them in weight and statistics spaces.

Model Connectivity and Post-Convergence States. Mode connectivity studies have revealed substantial geometric structure among neural-network solutions in parameter space. Independently trained solutions can often be connected through low-loss curves in practice (Garipov et al., 2018; Draxler et al., 2018). Subsequent work on linear mode connectivity further showed, in particular, that solutions obtained under different SGD noise realizations can become linearly connected during training (Frankle et al., 2020).

Parameter states accumulated during optimization have also been studied as a source of improved model solutions. Stochastic weight averaging averages checkpoints sampled along an SGD trajectory to obtain a single model (Izmailov et al., 2018), while exponential moving averages maintain a temporally smoothed parameter state throughout training (Morales-Brotons et al., 2024). These approaches indicate that useful parameter information can extend beyond the final raw checkpoint. TRAILS builds on this perspective by treating the raw-to-EMA segment as a post-convergence decoding space, searching for a favorable interior weight state and subsequently re-estimating sport-specific normalization statistics.

3 METHOD

As illustrated in Fig. 2, TRAILS performs post-convergence decoding in weight and statistics spaces. We first formulate the unified AQA setting, then introduce Trajectory Interpolation (TI) for selecting a shared weight configuration and Statistics Recalibration (SR) for re-estimating sport-specific normalization statistics, and finally describe the inference procedure.

3.1 UNIFIED AQA

Given an input video $\mathbf{x}$, action quality assessment predicts a continuous quality score

$$\hat{y} = F(\mathbf{x}; \mathbf{w}), \tag{1}$$

where $F(\cdot; \mathbf{w})$ denotes the AQA model parameterized by $\mathbf{w}$. Unified AQA extends this setting to K heterogeneous sports with training sets

$$\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{K}, \tag{2}$$

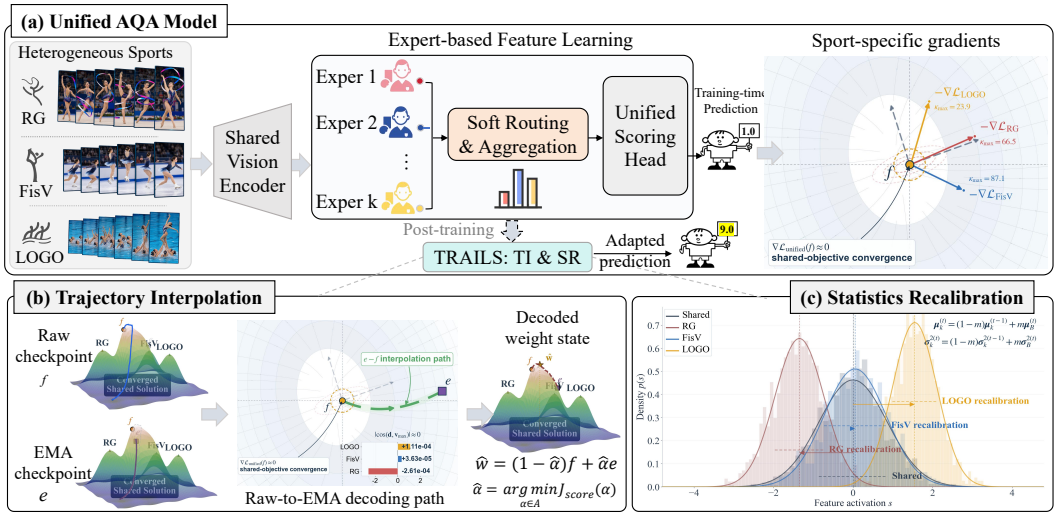


Figure 2: Overview of TRAILS. (a) Unified AQA learns a shared model across heterogeneous sports. (b) Trajectory Interpolation searches the raw-to-EMA segment for a favorable interior shared weight state $\hat{\mathbf{w}}$. (c) Statistics Recalibration re-estimates sport-specific BatchNorm statistics for the decoded weights, yielding sport-specific normalization states under the same shared weight configuration.

where each $\mathcal{D}_k$ contains video-score pairs $(\mathbf{x}, y)$ from sport k, and a single set of learnable parameters is shared across all sports.

Following MoAKE (Xu et al., 2026a), each video is represented by a sequence of 1024-dimensional visual features extracted by a pretrained Video Swin-T encoder (Liu et al., 2022) from 32-frame clips sampled at 25 fps. The resulting features are processed by action experts, while a soft router produces input-dependent expert weights for feature aggregation. A shared scoring head then predicts the normalized action-quality score. The complete MoAKE (Xu et al., 2026a) architecture is provided in the Appendix.

For sport k, the training objective is

$$\mathcal{L}_k(\mathbf{w}) = \lambda_1 \mathcal{L}_k^{\text{score}} + \lambda_2 \mathcal{L}_k^{\text{route}} + \lambda_3 \mathcal{L}_k^{\text{div}}, \tag{3}$$

where the three terms supervise quality prediction, soft routing, and expert diversity, respectively. Unified optimization aggregates the sport-specific objectives as

$$\mathcal{L}(\mathbf{w}) = \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\mathbf{w}), \tag{4}$$

where

$$\pi_k = \frac{N_k}{\sum_{j=1}^{K} N_j}, \tag{5}$$

and N_k denotes the number of training samples from sport k.

Unified optimization produces a shared parameter configuration shaped by heterogeneous sport-specific preferences, as characterized in Sec. 4.4. TRAILS subsequently revisits the model states produced during training and decodes them after convergence.

3.2 Trajectory Interpolation

Raw-to-EMA Trajectory. During unified training, let $\mathbf{w}^{(t)}$ denote the instantaneous model parameters at optimization step t. Alongside the raw model, an exponential moving average (EMA) companion is maintained as

$$\mathbf{w}_e^{(t)} = \beta \mathbf{w}_e^{(t-1)} + (1-\beta)\mathbf{w}^{(t)}, \tag{6}$$

where β denotes the EMA decay factor, set to 0.975 in our implementation. By exponentially weighting successive parameter states, EMA provides a temporally smoothed view of the recent training trajectory (Morales-Brotons et al., 2024).

After convergence, we denote the final raw checkpoint by $\mathbf{f}$ and its EMA companion by $\mathbf{e}$. Both originate from the same optimization trajectory but summarize its terminal dynamics differently. The raw checkpoint preserves the final instantaneous parameter state, whereas the EMA companion integrates recent parameter states through temporal smoothing. We use these two terminal views to define a linear post-convergence interpolation segment,

$$\mathbf{w}(\alpha) = (1-\alpha)\mathbf{f} + \alpha\mathbf{e}, \qquad \alpha \in [0,1], \tag{7}$$

where $\mathbf{w}(0) = \mathbf{f}$ and $\mathbf{w}(1) = \mathbf{e}$. The corresponding displacement is

$$\mathbf{d} = \mathbf{e} - \mathbf{f}. \tag{8}$$

We examine this segment as a post-convergence decoding space. As analyzed in Sec. 4.4, the training-side score objective remains low along the evaluated segment and attains its lowest evaluated value at an interior state. This observation motivates a one-dimensional search for a shared weight configuration between the raw checkpoint and its EMA companion.

Candidate Construction and Selection. Trajectory Interpolation searches the raw-to-EMA segment for a shared weight configuration. Given an interpolation grid $\mathcal{A} \subset [0,1]$, we construct

$$\mathcal{W}_{\mathrm{TI}} = \{\mathbf{w}(\alpha) \mid \alpha \in \mathcal{A}\}, \tag{9}$$

with

$$\mathcal{A} = \{0, 0.05, 0.10, \ldots, 0.95, 1.00\}. \tag{10}$$

For each candidate $\mathbf{w}(\alpha)$, we evaluate its training-side score objective on the unified training data:

$$\mathcal{J}_{\mathrm{score}}(\alpha) = \sum_{k=1}^{K} \pi_k \mathcal{L}_k^{\mathrm{score}}\left(\mathbf{w}(\alpha); \mathcal{D}_k\right), \tag{11}$$

where $\mathcal{L}_k^{\mathrm{score}}$ denotes the score prediction loss for sport k. The interpolation coefficient is selected according to

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \mathcal{J}_{\mathrm{score}}(\alpha). \tag{12}$$

Under the evaluated training configuration, the search selects $\hat{\alpha} = 0.50$, yielding the decoded shared weight configuration

$$\hat{\mathbf{w}} = \mathbf{w}(\hat{\alpha}). \tag{13}$$

3.3 Statistics Recalibration

Unified AQA uses a shared weight configuration, while heterogeneous sports can induce different activation distributions. BatchNorm running statistics accumulated during unified training reflect the mixed training distribution. Statistics Recalibration (SR) adapts these non-learnable statistics to each sport using unlabeled test samples, while keeping the decoded AQA weights fixed.

For sport k, we denote its recalibrated normalization state by

$$\mathcal{S}_k = \left\{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2\right\}, \tag{14}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$ collect the BatchNorm running means and variances, respectively.

Sport identities are inferred automatically using a lightweight classifier trained on frozen visual features from the training split. The unlabeled test set is grouped according to predicted sport identities. For each group, the decoded AQA model is forwarded with all learnable parameters fixed and dropout disabled, updating only the BatchNorm running statistics. No test quality scores are used for recalibration. At recalibration step t, the running mean is updated as

$$\boldsymbol{\mu}_k^{(t)} = (1-m)\boldsymbol{\mu}_k^{(t-1)} + m\boldsymbol{\mu}_B^{(t)}, \tag{15}$$

and the running variance as

$$\boldsymbol{\sigma}_k^{2(t)} = (1-m)\boldsymbol{\sigma}_k^{2(t-1)} + m\boldsymbol{\sigma}_B^{2(t)}, \tag{16}$$

where $\boldsymbol{\mu}_B^{(t)}$ and $\boldsymbol{\sigma}_B^{2(t)}$ denote the batch mean and variance, and m is the BatchNorm momentum.

Repeating this procedure for all sports yields the normalization states

$$\{\mathcal{S}_k\}_{k=1}^{K}. \tag{17}$$

TRAILS thus retains a single shared AQA weight configuration while maintaining sport-specific BatchNorm statistics estimated from unlabeled test data. These normalization states adapt the shared model to each sport without updating its learnable parameters.

3.4 Inference

After post-convergence decoding and test-time statistics recalibration, TRAILS retains the shared weights $\hat{\mathbf{w}}$ together with sport-specific normalization states $\{\mathcal{S}_k\}_{k=1}^{K}$. For an input video $\mathbf{x}$, its sport identity $\hat{k}$ is predicted from frozen visual features. The corresponding normalization state is then used to produce the final quality score:

$$\hat{y} = F\left(\mathbf{x}; \hat{\mathbf{w}}, \mathcal{S}_{\hat{k}}\right), \tag{18}$$

where $\hat{k}$ denotes the predicted sport identity.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate TRAILS under the unified AQA setting on three heterogeneous long-term benchmarks: RG (Zeng & Zheng, 2024), FisV (Xu et al., 2022a), and LOGO (Zhang et al., 2023). Following established AQA evaluation protocols (Ke et al., 2024; Yu et al., 2021; Zhou et al., 2025; Xu et al., 2026b), we report Spearman's rank correlation coefficient (SRCC) and relative ℓ_2 error (R-ℓ_2). Higher SRCC and lower R-ℓ_2 indicate better performance.

Implementation details. TRAILS is built on the unified MoAKE architecture (Xu et al., 2026a). Following its training protocol, we train the AQA model for 380 epochs using Adam with an initial learning rate of 3.4×10^{-4}, cosine learning-rate decay, and a batch size of 32. The EMA decay factor is set to 0.975. We denote the final raw checkpoint by $\mathbf{f}$ and its EMA companion by $\mathbf{e}$.

4.2 Comparison with State of the Art

Table 1 compares TRAILS with representative one-by-one and unified AQA methods across RG, FisV, and LOGO. Among one-by-one methods, CoFInAl (Zhou et al., 2024a) achieves the strongest average SRCC of 0.744, providing a strong reference for sport-specific training. However, when existing methods are extended to the unified setting, their performance generally degrades, reflecting the difficulty of preserving sport-specific performance within a shared model.

In contrast, TRAILS achieves an overall SRCC of 0.777 and an R-ℓ_2 of 2.008 under the unified AQA setting, obtaining the highest overall SRCC and lowest overall R-ℓ_2 among the evaluated unified methods. Compared with the reproduced MoAKE (Xu et al., 2026a), TRAILS improves SRCC from 0.756 to 0.778 on RG, from 0.659 to 0.673 on FisV, and from 0.837 to 0.850 on LOGO. The overall R-ℓ_2 also decreases from 2.074 to 2.008, yielding more accurate score predictions overall. These results demonstrate that post-convergence decoding can further improve the shared cross-sport scoring solution produced by unified training.

4.3 Ablation Studies

We examine the contributions of TI and SR, the selection of the interpolation coefficient, and sport-specific statistics recalibration.

Table 1: Quantitative comparison on three long-term AQA benchmarks under one-by-one and unified evaluation settings. **Bold** indicates the best result for each metric among the evaluated unified methods.

Type	Methods	RG		FisV		LOGO		Average	
		SRCC↑	R-ℓ_2↓	SRCC↑	R-ℓ_2↓	SRCC↑	R-ℓ_2↓	SRCC↑	R-ℓ_2↓
One by One	GDLT (Xu et al., 2022a) (CVPR'22)	0.765	2.401	0.685	3.717	0.647	4.148	0.703	3.422
	HGCN (Zhou et al., 2023) (TCSVT'23)	0.735	4.341	0.246	12.628	0.671	6.564	0.583	7.844
	T^{2}CR (Ke et al., 2024) (Inf. Sci.'24)	0.727	4.054	0.652	5.113	0.681	5.973	0.688	5.047
	PAMFN (Zeng & Zheng, 2024) (TIP'24)	0.711	4.561	0.665	7.235	0.659	13.979	0.679	8.592
	CoFInAl (Zhou et al., 2024a) (IJCAI'24)	**0.807**	4.028	0.716	2.875	0.698	4.019	0.744	3.641
	ASGTN (Liu et al., 2025) (TCSVT'25)	0.799	3.027	0.703	3.039	0.704	5.695	0.739	3.920
All in One	GDLT (Xu et al., 2022a) (CVPR'22)	0.702	2.716	0.377	4.681	0.447	5.499	0.525	4.299
	HGCN (Zhou et al., 2023) (TCSVT'23)	0.507	4.621	0.441	2.935	0.358	6.181	0.437	4.579
	T^{2}CR (Ke et al., 2024) (Inf. Sci.'24)	0.686	3.164	0.594	3.569	0.704	20.828	0.664	9.187
	PAMFN (Zeng & Zheng, 2024) (TIP'24)	0.683	3.399	0.502	2.849	0.690	13.082	0.632	6.443
	CoFInAl (Zhou et al., 2024a) (IJCAI'24)	0.741	2.834	0.496	3.032	0.603	6.580	0.624	4.149
	ASGTN (Liu et al., 2025) (TCSVT'25)	0.767	3.558	0.602	3.790	0.545	4.809	0.649	4.052
	MCMoE (Xu et al., 2026c) (AAAI'26)	0.771	3.426	0.625	3.608	0.763	4.816	0.726	3.950
	QGVL (Xu et al., 2025b) (TMM'25)	0.736	2.240	0.592	4.377	0.746	3.393	0.697	3.337
	MLAVL (Xu et al., 2025a) (CVPR'25)	0.755	2.655	0.632	2.919	0.765	5.043	0.722	3.539
	MoAKE (Xu et al., 2026a) (ECCV'26)	**0.801**	**1.822**	0.647	2.771	0.790	2.708	0.753	2.434
	MoAKE (Reproduced, **f**)	0.756	2.042	0.659	2.093	0.837	**2.085**	0.760	2.074
	TRAILS	0.778	1.826	**0.673**	**2.075**	**0.850**	2.123	**0.777**	**2.008**

Table 2: Component analysis of TRAILS.

Configuration	SRCC ↑	R-ℓ_2 ↓
EMA e	0.7512	2.109
Raw f	0.7600	2.074
f + SR	0.7683	2.096
TI, $\mathbf{w}(0.50)$	0.7651	**1.998**
TI + SR	**0.7768**	2.008

Table 3: Training-side score objective values.

α	$\mathcal{J}_{\text{score}}$
0.00	1.155259
0.30	1.112593
0.50	**1.102519**
0.70	1.110074
1.00	1.167407

Table 4: Effect of statistics recalibration strategies.

Configuration	SRCC ↑
TI without SR	0.7651
TI + pooled SR	0.7675
TI + oracle SR	**0.7768**
TI + automatic SR	**0.7768**
TI + imperfect-ID SR	0.7291

Component analysis. Table 2 presents the component ablation of TRAILS. Starting from the raw checkpoint f, TI improves the average SRCC from 0.7600 to 0.7651 and reduces R-ℓ_2 from 2.074 to 1.998, showing that interpolation alone yields a better shared weight state. Applying SR directly to f increases the average SRCC to 0.7683, indicating that recalibrating sport-specific normalization statistics also improves the raw model. Combining TI and SR further raises the average SRCC to 0.7768, the highest among all configurations, with an R-ℓ_2 of 2.008, confirming that both components contribute to post-convergence decoding.

Trajectory coefficient. Table 3 reports representative values of the training-side score objective along the raw-to-EMA trajectory. Starting from the raw checkpoint at $\alpha = 0$, the objective decreases from 1.1553 to 1.1126 at $\alpha = 0.30$ and reaches its lowest evaluated value of 1.1025 at $\alpha = 0.50$, before rising to 1.1101 at $\alpha = 0.70$ and 1.1674 at the EMA endpoint. The grid search thus selects an interior state rather than either endpoint. We use $\hat{\alpha} = 0.50$ in the remaining experiments under this training configuration. Full results over all 21 interpolation coefficients are reported in the Appendix.

Sport-specific recalibration. Table 4 compares statistics recalibration strategies applied to the decoded weight state. *Pooled SR* estimates a single BatchNorm state from all unlabeled test videos, whereas *oracle SR* estimates sport-specific states using ground-truth sport identities. *Automatic SR* instead uses identities predicted by the lightweight sport classifier. To examine sensitivity to identification errors, *imperfect-ID SR* introduces 10% sport-identity corruption.

Pooled SR increases SRCC from 0.7651 to 0.7675, whereas oracle and automatic SR both reach 0.7768 in this comparison. The sport classifier achieves 100% classification accuracy on the evaluated test split. With 10% identity corruption, SRCC decreases to 0.7291, below pooled SR and TI without SR. These results support sport-specific recalibration under accurate sport identification and show sensitivity to identity errors in the evaluated setting.

Table 5: Local optimization geometry at the raw checkpoint $\mathbf{f}$. Pairwise gradient cosine similarities characterize the agreement among sport-specific optimization signals. $\mathbf{d} = \mathbf{e} - \mathbf{f}$ denotes the raw-to-EMA displacement, and $\mathbf{v}_{\max}^{(k)}$ denotes the dominant Hessian eigendirection.

Sport	$\|\mathbf{g}_k\|_2$	$\cos(\mathbf{g}_k, \mathbf{g}_{\mathrm{RG}})$	$\cos(\mathbf{g}_k, \mathbf{g}_{\mathrm{FisV}})$	$\cos(\mathbf{g}_k, \mathbf{g}_{\mathrm{LOGO}})$	$\cos(\mathbf{g}_k, \mathbf{d})$	Dominant curvature	$\lvert\cos(\mathbf{v}_{\max}^{(k)}, \mathbf{d})\rvert$
RG	1.1349	1.000	0.090	-0.126	$+2.03 \times 10^{-4}$	113.863	1.93×10^{-4}
FisV	3.9046	0.090	1.000	-0.083	-8.780×10^{-4}	387.25	3.80×10^{-4}
LOGO	2.1609	-0.126	-0.083	1.000	$+1.07 \times 10^{-4}$	43.759	8.40×10^{-5}

4.4 Trajectory Analysis

We examine the raw-to-EMA trajectory through local optimization geometry, the distribution of parameter displacement, and functional localization.

Optimization geometry at the raw checkpoint. Table 5 characterizes the local first- and second-order optimization geometry at the raw checkpoint $\mathbf{f}$. Sport-specific gradients remain nonzero and exhibit limited directional agreement. RG and FisV have a small positive cosine similarity of 0.090, whereas RG–LOGO and FisV–LOGO yield negative similarities of -0.126 and -0.083, respectively. Dominant Hessian curvature also varies substantially across sports. These results reveal heterogeneous local optimization signals around the shared checkpoint. We further characterize the raw-to-EMA displacement $\mathbf{d} = \mathbf{e} - \mathbf{f}$ relative to the local optimization geometry of the shared model. The displacement exhibits numerically small cosine similarities with both sport-specific gradients and dominant Hessian eigendirections, indicating limited directional alignment with the measured first- and second-order quantities across sports. Together with the interior minimum identified by the training-side score objective (Table 3), these observations motivate examining the raw-to-EMA segment as a candidate decoding space for identifying favorable post-convergence weight states.

Layer-wise displacement. Figure 3 examines the distribution of the raw-to-EMA parameter displacement across model components, measured by each component's normalized squared ℓ_2 displacement norm. The prediction head accounts for 7.79% of total displacement energy, while the remaining 92.21% is distributed across upstream components. Thus, most of the measured displacement occurs outside the prediction head.

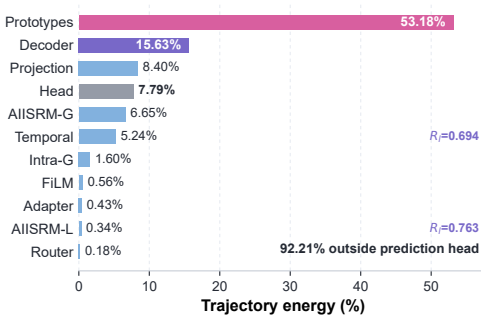


Figure 3: Layer-wise distribution of the raw-to-EMA displacement.

Figure 4 reports gradient–displacement alignment across model components. Absolute cosine similarities are below 0.008 for most components, with larger values in the Router and FiLM modules. The largest reported values are 0.064 for RG in the Router and 0.040 and 0.041 for FisV and LOGO in FiLM, respectively. These measurements describe variation in local alignment across functional components.

Functional localization. Table 6 examines TI by interpolating selected parameter groups at $\alpha = 0.50$, while retaining the raw checkpoint parameters elsewhere. Interpolating the prediction head alone or the late block together with the head does not improve average SRCC. Backbone-only interpolation increases average SRCC from 0.7600 to 0.7647, compared with 0.7651 under full interpolation. Most of the measured TI gain is therefore retained when interpolation is restricted to the backbone.

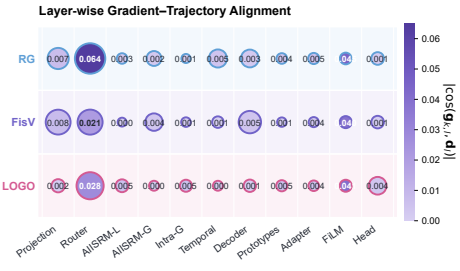


Figure 4: Layer-wise alignment between sport-specific gradients and the raw-to-EMA direction.

Table 7 compares TI with alternative post-convergence weight constructions. Late-checkpoint SWA obtains an average SRCC of 0.724–0.728, and the evaluated

Table 6: Functional localization of TI at $\alpha = 0.50$. Only the indicated parameter groups are interpolated.

Interpolated parameters	RG	FisV	LOGO	Average
Raw $\mathbf{f}$	0.7558	0.6586	0.8373	0.7600
Prediction head only	0.7542	0.6574	0.8373	0.7592
Late block + head	0.7536	0.6593	0.8353	0.7585
Backbone only	**0.7659**	**0.6737**	0.8320	0.7647
Full interpolation	0.7654	0.6729	**0.8336**	**0.7651**
EMA $\mathbf{e}$	0.7580	0.6600	0.8154	0.7512

Table 7: Alternative post-convergence weight constructions.

Configuration	SRCC ↑
Raw $\mathbf{f}$	0.7600
Late-checkpoint SWA	0.724–0.728
Independent interpolation	0.2331
Raw-to-EMA TI, $\mathbf{w}(0.50)$	**0.7651**
TRAILS, TI + SR	0.7768

interpolation between independently trained models obtains 0.2331. By comparison, raw-to-EMA TI increases average SRCC from 0.7600 to 0.7651. Under the evaluated configurations, the raw-to-EMA construction provides a more favorable weight state than these alternatives.

5 Qualitative Analysis

Figure 5 visualizes the representations immediately before the prediction head along the raw-to-EMA trajectory. Panels (a)–(c) show the embeddings of all three sports under the raw checkpoint f, EMA companion e, and decoded state $\mathbf{w}(\hat{\alpha})$, respectively. The three states exhibit noticeably different embedding structures, while the decoded state shows compact within-sport organization and clear separation across sports. Panel (d) further visualizes the displacement of RG representations along the trajectory, showing that the decoded features occupy an intermediate but distinct configuration between the raw and EMA states.

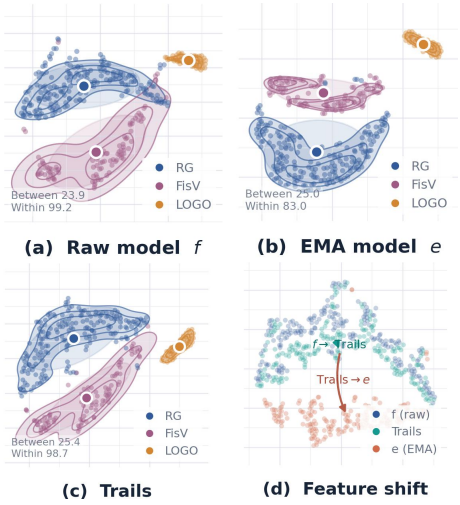


Figure 5: t-SNE visualization along the raw-to-EMA trajectory.

Figure 6 compares the reproduced MoAKE (Xu et al., 2026a) and TRAILS on four representative action sequences from different sports. Across the displayed examples, TRAILS produces predictions closer to the corresponding ground-truth scores, illustrating its improved score estimation at the individual sample level. Additional qualitative results across the three sports are provided in the Appendix.

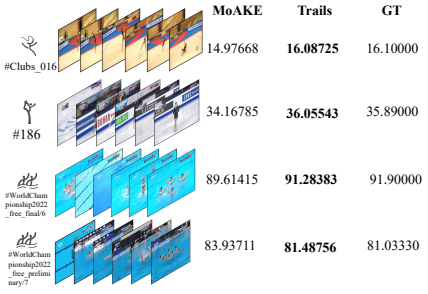


Figure 6: Qualitative score comparison.

6 Conclusion

We presented **TRAILS**, a post-convergence decoding framework for unified action quality assessment. Our analysis revealed limited directional agreement among sport-specific gradients, substantial variation in dominant local curvature, and limited alignment between the raw-to-EMA displacement and sport-specific gradients and dominant Hessian eigendirections. Building on these observations, TRAILS decodes the final raw checkpoint and its EMA companion in both weight and statistics spaces. Trajectory Interpolation identifies a favorable interior weight state along the raw-to-EMA path, while Statistics Recalibration re-estimates sport-specific BatchNorm statistics for the decoded model. The framework requires neither additional gradient-based retraining of the AQA model nor architectural modifications. Experiments on three heterogeneous AQA benchmarks show that TRAILS achieves an average SRCC of 0.777 and an R-ℓ_2 of 2.008, with the highest average SRCC and lowest average R-ℓ_2 among the evaluated unified methods.

AI USE STATEMENT

Generative AI tools, including ChatGPT and DeepSeek V4, were used to assist with language refinement and improve the clarity and readability of the manuscript. The tools were not used to generate experimental results or scientific claims. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work does not involve the collection of new human-subject data or the release of new datasets. All experiments are conducted on existing action quality assessment benchmarks and focus on model training, post-convergence analysis, and evaluation. We do not identify ethical risks beyond those associated with the use of the underlying benchmark data and models. The authors have reviewed the experimental design and reporting to ensure that the results and conclusions are presented accurately and responsibly.

REPRODUCIBILITY STATEMENT

We provide the main implementation and experimental details required to reproduce TRAILS in the paper and Appendix. The model architecture, unified training configuration, optimization settings, EMA construction, interpolation grid, coefficient-selection objective, and Statistics Recalibration procedure are described in detail. The Appendix further reports the complete trajectory coefficient search, sport-specific optimization geometry, recalibration settings, perturbation analysis, and additional qualitative results. These details are intended to support independent reproduction of the reported results.

REFERENCES

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.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International conference on machine learning*, pp. 3259–3269. PMLR, 2020.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Fine-grained spatio-temporal parsing network for action quality assessment. *IEEE Transactions on Image Processing*, 32: 6386–6400, 2023.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

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

Xiao Ke, Huangbiao Xu, Xiaofeng Lin, and Wenzhong Guo. Two-path target-aware contrastive regression for action quality assessment. *Information Sciences*, 664:120347, 2024.

Mingzhe Li, Hong-Bo Zhang, Qing Lei, Zongwen Fan, Jinghua Liu, and Ji-Xiang Du. Pairwise contrastive learning network for action quality assessment. In *European conference on computer vision*, pp. 457–473. Springer, 2022.

Yicong Li and Howard Leung. Finegrade: A rule-consistent scoring framework for fine-grained action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 8258–8267, 2026.

Yuan-Ming Li, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Continual action assessment via task-consistent score-discriminative feature distribution modeling. *IEEE transactions on circuits and systems for video technology*, 34(10):9112–9124, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021a.

Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9517–9526. IEEE, 2021b.

Jiang Liu, Huasheng Wang, Wei Zhou, Katarzyna Stawarz, Padraig Corcoran, Ying Chen, and Hantao Liu. Adaptive spatiotemporal graph transformer network for action quality assessment. *IEEE transactions on circuits and systems for video technology*, 35(7):6628–6639, 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3192–3201. IEEE, 2022.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *arXiv preprint arXiv:2411.18704*, 2024.

Lauren Okamoto and Paritosh Parmar. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3204–3213, 2024.

Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6330–6339. IEEE, 2019.

Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1468–1476. IEEE, 2019a.

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

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9836–9845. IEEE, 2020.

Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2901–2909, 2023.

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

Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23967–23977. IEEE, 2025a.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Yuezhou Li, Rui Xu, and Wenzhong Guo. Quality-guided vision-language learning for long-term action quality assessment. *IEEE Transactions on Multimedia*, 2025b.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Jiaxin Cai, Junyi Wu, and Jinglin Xu. Moake: Toward unified all-in-one action quality assessment via mixture of action knowledge experts. In *European Conference on Computer Vision*, pp. 294–312. Springer, 2026a.

Huangbiao Xu, Huanqi Wu, Xiao Ke, and Yuxin Peng. Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. *arXiv preprint arXiv:2605.00434*, 2026b.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Junyi Wu, Rui Xu, and Jinglin Xu. Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11241–11249, 2026c.

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*, 2022b.

Jinglin Xu, Yongming Rao, Jie Zhou, and Jiwen Lu. Procedure-aware action quality assessment: Datasets and performance evaluation. *International journal of computer vision*, 132(12):6069–6090, 2024a.

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, 2024b.

Hui Yu, Xiao Ke, Zhihong Zeng, Huangbiao Xu, and Huanqi Wu. Harp: Hierarchical adaptive ranking with probabilistic modeling for skill determination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8337–8346, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

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

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

Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2405–2414. IEEE, 2023.

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

Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. *IJCAI*, 2024a.

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*, pp. 375–392. Springer, 2024b.

Kanglei Zhou, Hubert PH Shum, Frederick WB Li, Xingxing Zhang, and Xiaohui Liang. Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE transactions on image processing*, 34:3718–3732, 2025.

Appendix–LEARNING FROM WHAT IT LEARNED: DECODING POST-CONVERGENCE STATES FOR UNIFIED ACTION QUALITY ASSESSMENT

This appendix provides additional implementation details and experimental analyses that complement the main paper. We first describe the MoAKE baseline, unified training configuration, and model states used by TRAILS. We then visualize cross-sport heterogeneity in the shared visual feature space and examine sport-specific first- and second-order optimization geometry, including random-direction controls. Next, we report the complete trajectory coefficient search and additional sensitivity analyses of EMA decay and local weight perturbations. Finally, we provide further qualitative score comparisons across RG, FisV, and LOGO.

A Baseline Architecture and Experimental Setup

A.1 MoAKE Architecture

TRAILS is built upon MoAKE (Xu et al., 2026a), a unified action quality assessment framework that learns a shared model across heterogeneous sports. Figure 7 illustrates the complete baseline architecture used in our experiments.

Given an input video, a pretrained Video Swin-T encoder extracts 1024-dimensional visual features from 32-frame clips sampled at 25 fps. The resulting features are processed by Action Knowledge Experts. As shown in Fig. 7(b), each expert contains token projection, temporal context enhancement, a shared adapter, segment-aware temporal aggregation, and Adaptive Intra- and Inter-Segment Relationship Modeling (AIISRM). These components transform the shared visual features into expert-specific temporal representations.

A soft router predicts input-dependent expert weights and aggregates the expert representations into a mixed multi-expert representation. AIISRM models intra-segment global, inter-segment local, and inter-segment global relationships, as illustrated in Fig. 7(c). The mixed expert representation is then processed by a grade-aware decoder and an MLP to produce the final quality prediction. The MoAKE training objective jointly supervises score prediction, expert routing, and expert diversity.

TRAILS retains the complete MoAKE architecture and operates on model states obtained after unified training.

A.2 Unified Training Setup

The shared model is trained for 380 epochs using Adam with an initial learning rate of 3.4×10^{-4}, cosine learning-rate decay, and a batch size of 32. The random seed is fixed to 109. An exponential moving average (EMA) model is maintained throughout training with decay factor

$$\beta = 0.975. \tag{19}$$

For sport k, the unified model optimizes

$$\mathcal{L}_k(\mathbf{w}) = \lambda_1 \mathcal{L}_k^{\text{score}} + \lambda_2 \mathcal{L}_k^{\text{route}} + \lambda_3 \mathcal{L}_k^{\text{div}}, \tag{20}$$

where the three terms supervise quality prediction, expert routing, and expert diversity, respectively.

The shared training objective is

$$\mathcal{L}(\mathbf{w}) = \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\mathbf{w}), \qquad \pi_k = \frac{N_k}{\sum_{j=1}^{K} N_j}, \tag{21}$$

where N_k denotes the number of training samples from sport k.

A.3 Raw and EMA States

During unified training, the instantaneous parameters $\mathbf{w}^{(t)}$ are updated by gradient-based optimization, while an EMA companion is maintained as

$$\mathbf{w}_e^{(t)} = \beta \mathbf{w}_e^{(t-1)} + (1-\beta)\mathbf{w}^{(t)}. \tag{22}$$

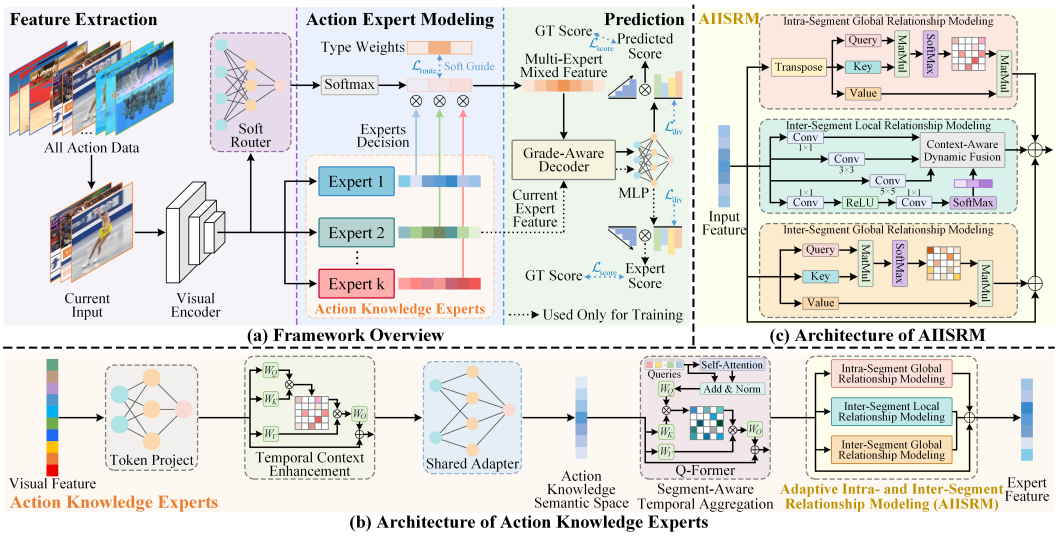


Figure 7: Architecture of the MoAKE baseline used in our experiments. The framework consists of feature extraction, action expert modeling, and score prediction. Action Knowledge Experts learn expert-specific temporal representations, while a soft router produces input-dependent expert weights. AIISRM models intra-segment global, inter-segment local, and inter-segment global relationships. TRAILS retains the complete MoAKE architecture and decodes model states from the unified training trajectory.

The EMA parameters do not directly participate in gradient updates. Instead, exponential averaging aggregates recent parameter states and provides a temporally smoothed view of the training trajectory (Morales-Brotons et al., 2024).

After convergence, we denote the final raw checkpoint by $\mathbf{f}$ and its EMA companion by $\mathbf{e}$. Both originate from the same optimization trajectory but summarize its terminal states differently: $\mathbf{f}$ is the final instantaneous parameter state, whereas $\mathbf{e}$ incorporates recent states through exponential averaging.

A.4 Sport Identification and Statistics Recalibration

Sport identities are predicted using a lightweight linear classifier trained on frozen 1024-dimensional Video Swin-T features from the training split. The classifier operates on pooled visual features and achieves 100% sport classification accuracy on the evaluated test split.

For automatic Statistics Recalibration (SR), the unlabeled test set is grouped according to predicted sport identities. The decoded AQA model is forwarded over each group to estimate its sport-specific BatchNorm statistics, with learnable AQA parameters fixed and dropout disabled. The resulting statistics are used when evaluating samples from the corresponding sport.

This procedure uses the unlabeled test set collectively for transductive recalibration. Neither sport classification nor BatchNorm recalibration uses test quality scores.

B Datasets and Cross-Sport Heterogeneity

We evaluate unified AQA on RG (Zeng & Zheng, 2024), FisV (Xu et al., 2022a), and LOGO (Zhang et al., 2023). These benchmarks differ in motion patterns, temporal organization, action duration, object interaction, camera characteristics, and score distributions. RG additionally contains four apparatus categories: Ball, Clubs, Hoop, and Ribbon. Despite these differences, samples from all three benchmarks are processed by the same visual encoder and optimized within a single shared AQA model.

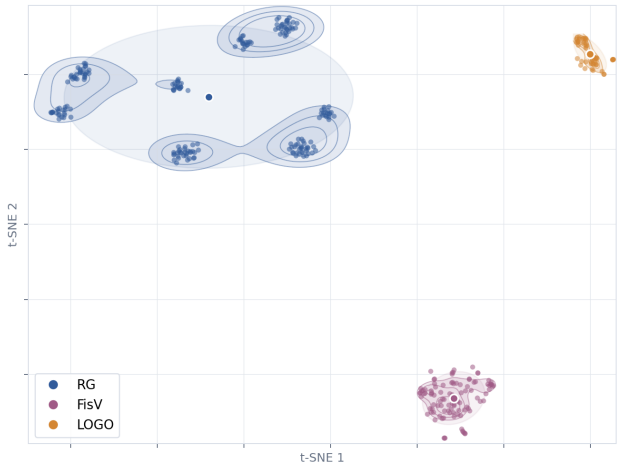


Figure 8: t-SNE visualization of pretrained Video Swin-T features extracted from RG, FisV, and LOGO.

B.1 Visual Feature Heterogeneity

Figure 8 visualizes features extracted by the pretrained Video Swin-T encoder before unified AQA modeling. Samples from RG, FisV, and LOGO occupy distinguishable regions in the displayed t-SNE embedding, illustrating differences in their visual representations.

This visualization provides a representation-level view of the heterogeneous inputs encountered by the shared model. We next examine the corresponding sport-specific optimization signals around the converged checkpoint.

C Sport-Specific Optimization Geometry

We examine the local optimization geometry around the converged shared model. The measurements in Table 8 are performed at the raw checkpoint $\mathbf{f}$ using the complete training sets of RG, FisV, and LOGO, containing 800, 400, and 150 samples, respectively. Deterministic center cropping is used throughout. Gradients and Hessian-vector products are computed using the complete MoAKE training objective over all 13.9 million trainable parameters.

C.1 Sport-Specific First-Order Geometry

For sport k, let $\mathcal{L}_k(\mathbf{w})$ denote its training objective. The full-training-set gradient at the raw checkpoint is

$$\mathbf{g}_k = \nabla_{\mathbf{w}} \mathcal{L}_k(\mathbf{f}). \tag{23}$$

We measure directional agreement between two sport-specific gradients by

$$C_{ij} = \cos(\mathbf{g}_i, \mathbf{g}_j) = \frac{\mathbf{g}_i^\top \mathbf{g}_j}{\|\mathbf{g}_i\|_2 \|\mathbf{g}_j\|_2}. \tag{24}$$

The raw-to-EMA displacement is defined as

$$\mathbf{d} = \mathbf{e} - \mathbf{f}, \tag{25}$$

and its first-order alignment with sport k is measured by

$$A_k^{(1)} = \cos(\mathbf{g}_k, \mathbf{d}). \tag{26}$$

Table 8: Full-training-set first- and second-order optimization geometry at the converged raw checkpoint $\mathbf{f}$. Gradients and Hessian-vector products are computed using the complete MoAKE training objective. Dominant curvature is estimated by 25-step power iteration. The raw-to-EMA displacement is $\mathbf{d} = \mathbf{e} - \mathbf{f}$.

Sport	$\|\mathbf{g}_k\|_2$	$\cos(\mathbf{g}_k, \mathbf{g}_{\mathrm{RG}})$	$\cos(\mathbf{g}_k, \mathbf{g}_{\mathrm{FisV}})$	$\cos(\mathbf{g}_k, \mathbf{g}_{\mathrm{LOGO}})$	$\cos(\mathbf{g}_k, \mathbf{d})$	$\lambda^{(k)}_{\mathrm{dom}}$	$\lvert\cos(\mathbf{v}^{(k)}_{\mathrm{dom}}, \mathbf{d})\rvert$	$\lvert\cos(\mathbf{v}^{(k)}_{\mathrm{dom}}, \mathbf{g}_k)\rvert$
RG	1.134879	1.000000	0.089806	-0.125986	$+2.03 \times 10^{-4}$	113.8630	1.93×10^{-4}	0.981807
FisV	3.904559	0.089806	1.000000	-0.082687	-8.780×10^{-4}	387.25	3.80×10^{-4}	0.929430
LOGO	2.160896	-0.125986	-0.082687	1.000000	$+1.07 \times 10^{-4}$	43.7587	8.40×10^{-5}	0.915772

C.2 Dominant Second-Order Geometry

For sport k, the Hessian of its training objective at $\mathbf{f}$ is

$$\mathbf{H}_k = \nabla^2_{\mathbf{w}} \mathcal{L}_k(\mathbf{f}). \tag{27}$$

Explicitly constructing the Hessian is impractical at this parameter scale. We therefore compute Hessian-vector products through second-order automatic differentiation. For a vector $\mathbf{v}$,

$$\mathbf{H}_k \mathbf{v} = \nabla_{\mathbf{w}} \left(\mathbf{g}_k^\top \mathbf{v} \right). \tag{28}$$

Starting from a normalized initial vector, we apply 25 steps of power iteration,

$$\mathbf{v}^{(t+1)} = \frac{\mathbf{H}_k \mathbf{v}^{(t)}}{\|\mathbf{H}_k \mathbf{v}^{(t)}\|_2}, \tag{29}$$

to estimate the dominant eigendirection $\mathbf{v}^{(k)}_{\mathrm{dom}}$. The corresponding curvature is estimated using the Rayleigh quotient

$$\lambda^{(k)}_{\mathrm{dom}} = \left(\mathbf{v}^{(k)}_{\mathrm{dom}} \right)^\top \mathbf{H}_k \mathbf{v}^{(k)}_{\mathrm{dom}}. \tag{30}$$

Because the sign of an eigenvector is arbitrary, comparisons involving Hessian eigendirections use absolute cosine similarity.

Table 8 reports the local optimization geometry of the three sports. Their gradients remain nonzero after unified training and exhibit limited directional agreement. The cosine similarity is 0.090 for RG–FisV, -0.126 for RG–LOGO, and -0.083 for FisV–LOGO. These measurements show that sport-specific optimization signals persist around the shared checkpoint.

The raw-to-EMA displacement exhibits small numerical cosine similarities with individual sport gradients, with absolute values below 10^{-3}. Dominant curvature also varies substantially across sports, taking values of 113.863, 387.25, and 43.759 for RG, FisV, and LOGO, respectively. The corresponding dominant Hessian eigendirections exhibit absolute cosine similarities below 4×10^{-4} with the raw-to-EMA displacement.

The dominant eigendirections are also closely aligned with their corresponding sport gradients, with absolute cosine similarities of 0.982, 0.929, and 0.916. Together, these measurements characterize heterogeneous sport-specific local geometry and the directional relationships among gradients, dominant Hessian eigendirections, and the raw-to-EMA displacement.

C.3 Random-Direction Controls

We further contextualize the raw-to-EMA displacement using isotropically sampled random directions in the $13{,}904{,}924$-dimensional parameter space. For first-order alignment, we evaluate 512 random directions and compare their absolute gradient cosine similarities with those of the raw-to-EMA displacement.

For second-order geometry, we evaluate directional curvature using the Rayleigh quotient

$$q_k(\mathbf{u}) = \frac{\mathbf{u}^\top \mathbf{H}_k \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}, \tag{31}$$

Table 9: Random-direction controls for gradient alignment and directional curvature. The first-order comparison uses 512 isotropic random directions; the second-order comparison uses 32. Intervals denote empirical 95% ranges over the sampled random directions.

Sport	Absolute gradient cosine similarity			Directional Rayleigh quotient		
	Raw-to-EMA	Random mean	Random 95% interval	Raw-to-EMA	Random mean	Random 95% interval
RG	2.030×10^{-4}	2.153×10^{-4}	$[1.069 \times 10^{-5}, 5.840 \times 10^{-4}]$	$+7.026 \times 10^{-5}$	2.073×10^{-5}	$[6.548 \times 10^{-6}, 4.464 \times 10^{-5}]$
FisV	8.780×10^{-4}	2.059×10^{-4}	$[9.433 \times 10^{-6}, 5.777 \times 10^{-4}]$	$+1.940 \times 10^{-4}$	4.680×10^{-5}	$[7.569 \times 10^{-6}, 1.797 \times 10^{-4}]$
LOGO	1.070×10^{-4}	2.171×10^{-4}	$[2.339 \times 10^{-5}, 5.729 \times 10^{-4}]$	-1.676×10^{-4}	6.512×10^{-6}	$[6.161 \times 10^{-7}, 1.763 \times 10^{-5}]$

Table 10: Training-side score objective values over the complete 21-point raw-to-EMA interpolation grid.

α	Score Obj.	α	Score Obj.	α	Score Obj.	α	Score Obj.
0.00	1.155259	0.30	1.112593	0.60	1.103852	0.90	1.141926
0.05	1.146000	0.35	1.108444	0.65	1.106296	0.95	1.154074
0.10	1.137778	0.40	1.105185	0.70	1.110074	1.00	1.167407
0.15	1.130667	0.45	1.103407	0.75	1.115704		
0.20	1.123852	**0.50**	**1.102519**	0.80	1.122667		
0.25	1.117778	0.55	1.102667	0.85	1.131407		

where $\mathbf{u}$ denotes a nonzero parameter-space direction. We compare the raw-to-EMA direction with 32 isotropic random directions for each sport.

Table 9 places the numerical gradient alignments in a high-dimensional reference distribution. For RG and LOGO, the displacement alignments fall within the empirical 95% intervals of isotropic random directions. The FisV alignment exceeds the corresponding sampled interval, indicating variation in this relationship across sports.

The directional Rayleigh quotients provide an additional measurement of curvature along the raw-to-EMA direction. Their values are substantially smaller in magnitude than the dominant curvature values reported in the local geometry analysis, while their positions relative to the random-direction distributions vary across sports. These controls contextualize the directional measurements used to characterize the terminal training trajectory.

D Additional Ablation Studies

We provide additional analyses of trajectory coefficient selection, random perturbation paths, sensitivity to the EMA decay factor, and the robustness of the decoded weight state to local parameter perturbations.

D.1 Complete Trajectory Coefficient Search

Trajectory Interpolation evaluates the uniformly spaced coefficient set

$$\mathcal{A} = \{0, 0.05, 0.10, \ldots, 0.95, 1.00\}. \tag{32}$$

For each candidate $\mathbf{w}(\alpha)$, the training-side score objective $\mathcal{J}_{\text{score}}(\alpha)$ is evaluated using the unified training data. Table 10 reports the complete 21-point search. The objective decreases from the raw endpoint, reaches its lowest evaluated value at $\alpha = 0.50$, and increases toward the EMA endpoint.

The neighboring coefficients $\alpha = 0.45$ and $\alpha = 0.55$ yield similarly low objective values, showing that the selected coefficient lies within an interior region of low training-side score objective. We therefore use $\hat{\alpha} = 0.50$ for the subsequent experiments under this training configuration.

D.2 Random Perturbation Paths

We examine whether moving away from the raw checkpoint along random parameter-space directions reproduces the gains obtained by TRAILS. For each perturbation level, we sample an isotropic

Table 11: SRCC along random perturbation paths at different relative displacement magnitudes, compared with the raw-to-EMA path. Retrospective Peak SRCC denotes the highest observed value over the evaluated grid and is reported for retrospective analysis; TRAILS selects $\hat{\alpha} = 0.50$ using the training-side score objective.

Configuration	$\alpha = 0$	Retrospective Peak SRCC	$\alpha = 0.50$	$\alpha = 1$
Random, 1%	0.7683	0.7685	0.7685	0.7680
Random, 5%	0.7682	0.7690	0.7672	0.7675
Random, 10%	0.7683	0.7692	0.7674	0.7671
Raw-to-EMA + SR	0.7683	**0.7794**	**0.7768**	0.7589

Table 12: SRCC sensitivity of trajectory decoding to the EMA decay factor.

EMA decay	Raw $\mathbf{f}$	EMA $\mathbf{e}$	TI $\mathbf{w}(0.50)$	TRAILS
$\beta = 0.975$	0.7600	0.7512	0.7651	**0.7768**
$\beta = 0.99$	0.7600	0.7449	0.7680	0.7727
$\beta = 0.999$	0.7600	0.3675	0.7409	0.7406

random direction $\mathbf{r}$ over the trainable parameters and construct

$$\mathbf{w}_{\mathrm{rand}}(\alpha) = \mathbf{f} + \alpha\rho\|\mathbf{f}\|_2 \frac{\mathbf{r}}{\|\mathbf{r}\|_2}, \qquad \rho \in \{0.01, 0.05, 0.10\}, \tag{33}$$

where $\alpha \in \{0, 0.05, \ldots, 1\}$. Non-trainable model states are unchanged when constructing the perturbed weights. Each candidate is evaluated using the same Statistics Recalibration procedure.

Table 11 compares the raw-to-EMA path with random perturbation paths at three relative displacement magnitudes. The 1% random path exhibits little variation in SRCC, while the 5% and 10% paths yield modest improvements at some interior coefficients but generally decline at larger interpolation coefficients. The peak SRCC of each evaluated random path remains below that of the raw-to-EMA path.

At the training-side-selected coefficient $\hat{\alpha} = 0.50$, TRAILS achieves an SRCC of 0.7768, compared with 0.7685, 0.7672, and 0.7674 for the 1%, 5%, and 10% random paths, respectively. These results show that the evaluated fixed-relative-magnitude random perturbations do not reproduce the decoding gain obtained along the raw-to-EMA trajectory.

D.3 Sensitivity to EMA Decay

Table 12 examines the sensitivity of trajectory decoding to the EMA decay factor. Under the default setting $\beta = 0.975$, TRAILS achieves an SRCC of 0.7768, the highest among the three evaluated decay values.

At $\beta = 0.99$, TI improves over both the raw checkpoint and the EMA companion, and TRAILS reaches 0.7727. At $\beta = 0.999$, the EMA companion exhibits a substantial performance decrease, and neither TI nor TRAILS improves over the raw checkpoint. These results show that the evaluated decoding performance is sensitive to the EMA decay factor.

D.4 Local Weight Perturbation

We examine local neighborhoods of the raw checkpoint $\mathbf{f}$ and the decoded state $\mathbf{w}(0.50)$. Relative Gaussian perturbations are independently applied to each floating-point parameter tensor, with noise magnitude

$$\sigma \in \{0, 0.01, 0.02, 0.05, 0.10\}. \tag{34}$$

For each nonzero noise level, five independent perturbation runs are evaluated. A run is regarded as collapsed when the model produces near-constant predictions such that SRCC becomes undefined.

Table 13: Weight-space stability under relative Gaussian perturbations. SRCC statistics are computed over non-collapsed runs. Valid runs denotes the number of perturbation runs that preserve non-collapsed predictions.

State	Noise	SRCC	Valid runs
Raw $\mathbf{f}$	0.00	0.7600	5/5
	0.01	0.7593 ± 0.0006	5/5
	0.02	0.7593 ± 0.0010	2/5
	0.05	Collapse	0/5
	0.10	Collapse	0/5
$\mathbf{w}(0.50)$	0.00	0.7651	5/5
	0.01	0.7655 ± 0.0008	5/5
	0.02	0.7653 ± 0.0007	5/5
	0.05	0.7656 ± 0.0010	5/5
	0.10	0.7677 ± 0.0024	4/5

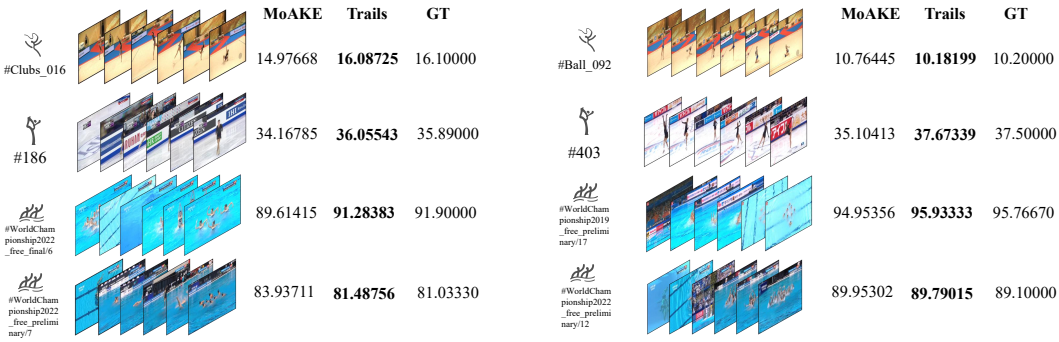


Figure 9: Additional qualitative comparison of score predictions produced by the reproduced MoAKE baseline and TRAILS on representative samples from RG, FisV, and LOGO. Across the displayed examples, TRAILS yields predictions closer to the corresponding ground-truth quality scores.

Table 13 compares the response of the raw and decoded states to relative Gaussian parameter perturbations. At $\sigma = 0.02$, the raw checkpoint produces non-collapsed predictions in two of five runs. All five raw-checkpoint runs collapse at $\sigma \geq 0.05$.

In contrast, the decoded state preserves valid predictions in all five runs up to $\sigma = 0.05$ and in four of five runs at $\sigma = 0.10$. Among non-collapsed runs, its SRCC remains close to the unperturbed performance across the evaluated noise levels. The results show greater resistance to perturbation-induced prediction collapse for the decoded state under this test.

E Additional Qualitative Results

E.1 Additional Score Prediction Visualizations

Figure 9 provides additional qualitative comparisons of TRAILS and the reproduced MoAKE baseline across RG, FisV, and LOGO. The displayed samples cover different visual content and action-quality levels. Across these examples, TRAILS produces predictions closer to the corresponding ground-truth scores, complementing the quantitative results in the main paper.