A^{2}QA: ANYTIME ACTION QUALITY ASSESSMENT VIA FLOW-BASED SCORE FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Action Quality Assessment (AQA) typically predicts a final performance score only after observing the complete video. However, in sports and skill training, predicting performance outcomes *before an action concludes* can enable earlier and more prospective feedback. We study the **anytime AQA** task in this paper, where the goal is to predict the final score from a partially observed performance while accounting for uncertainty in its unseen future. We propose **A^{2}QA**, a model-agnostic framework that augments a pretrained AQA model with plausible future completions. Given a partial video, A^{2}QA retrieves real future segments from training videos with compatible declared content and appends them to the partial video in feature space. To model uncertainty over possible futures, we introduce a conditional flow model that estimates the likelihood of each retrieved completion conditioned on the partial video. A frozen scoring model then evaluates each completed sequence, and their likelihood-weighted scores are aggregated to predict the final performance score. The framework requires neither retraining nor modification of the underlying scoring model and naturally reduces to the original model when the full performance is observed. Experiments on four datasets spanning sports and surgical skill assessment (FineFS, FineDiving, JIGSAWS, and LapEx) show consistent improvements across all 13 model. For example, in TSMP on FineFS datasets, Spearman correlation improves from 0.395 to 0.661 at only 10% observation. These results demonstrate that explicitly modeling plausible future completions substantially improves final-score prediction from partial observations while preserving compatibility with existing AQA models.

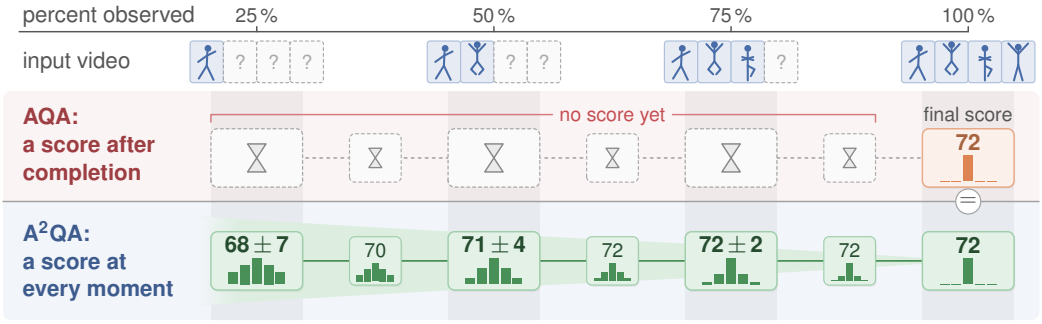

Figure 1: **Illustration of anytime AQA.** Conventional AQA scores complete routines (top); A^{2}QA predicts the final score from partial observations using retrieved endings (bottom).

1 INTRODUCTION

Action quality assessment (AQA) estimates how well an action is performed from video, with applications in sports analysis and skill assessment. Recent methods compare performances with reference examples (Yu et al., 2021), model action procedures (Xu et al., 2022b; 2024), or incorporate scoring rubrics (Majeedi et al., 2024). Existing methods operate under a *complete-observation* as-

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.

sumption: the entire action sequence is available before the final score is predicted. While appropriate for retrospective evaluation, this setting limits the ability of AQA systems to provide *prospective feedback* during an ongoing performance.

Anytime AQA is fundamentally different from conventional AQA because the final score depends not only on what has already occurred, but also on what may happen in the unseen future. A straightforward solution is to directly apply an existing AQA model to the truncated video. However, this creates a mismatch between training and inference: models trained on complete actions are asked to score incomplete sequences lacking potentially decisive events. More importantly, a single partial video can admit multiple plausible continuations with substantially different outcomes. For example, two performances may appear similar during their early stages but diverge later because of differences in execution, errors, or completion quality. Predicting a final score directly from the partial video therefore requires reasoning under uncertainty about these possible futures.

A natural way to address this uncertainty is to explicitly consider how the observed performance could continue. Rather than requiring the scoring model itself to extrapolate from an incomplete sequence, we ask a different question: *what plausible complete performances are consistent with the portion observed so far?* If such possible futures can be constructed, a conventional AQA model can continue operating in the setting for which it was originally designed by scoring complete action sequences. This perspective has two desirable properties. First, it decouples future reasoning from score prediction, allowing existing pretrained AQA models to be reused without redesigning or retraining them. Second, it provides a natural mechanism for representing the inherent multimodality of future actions by considering multiple possible completions rather than committing to a single predicted future.

Based on this idea, we propose **A^{2}QA**, a model-agnostic framework for anytime action quality assessment through plausible future completion. Given a partial video, A^{2}QA retrieves real future segments from training performances whose action or declared content is compatible with the current performance and attaches them to the partial video in feature space. Using real observed endings avoids the need to synthesize future video while preserving realistic motion and performance characteristics. The resulting sequences represent multiple hypotheses about how the ongoing action could ultimately unfold. Not all retrieved futures, however, are equally compatible with the observed performance. Simply averaging their predicted scores would ignore the relationship between what has already been observed and what is likely to occur next. We therefore introduce a conditional flow model that estimates the probability density of each candidate completion conditioned on the partial video. Each completed sequence is evaluated by the original frozen AQA model, and the resulting scores are aggregated according to their conditional likelihoods. In this way, A^{2}QA converts prediction from an incomplete action into an expectation over plausible completed actions. As the observation progresses, the space of plausible futures becomes increasingly constrained; once the full action is observed, no completion is necessary and the framework naturally recovers the prediction of the original scoring model.

We evaluate A^{2}QA on four benchmarks spanning sports and surgical skill assessment: FineFS, FineDiving, JIGSAWS, and LapEx. Across 13 combinations of AQA models and datasets, A^{2}QA consistently improves Spearman correlation at 10%, 30%, and 50% observation compared with directly applying the same frozen models to truncated videos. The advantage is particularly pronounced under very limited observation. These results demonstrate that explicitly reasoning over plausible future completions provides a simple yet effective mechanism for extending complete-video AQA models to prospective performance assessment.

Our main contributions are summarized as follows:

- We introduce the novel *anytime action quality assessment* task, which aims to predict the final performance score from partially observed actions while explicitly accounting for uncertainty in the unseen future.
- We propose A^{2}QA, a model-agnostic framework that transforms partial-observation AQA into complete-sequence scoring by retrieving plausible future segments from existing training performances, enabling pretrained AQA models to be reused without retraining.
- We introduce conditional density modeling over candidate futures, allowing the framework to weight multiple possible completions according to their compatibility with the observed action rather than relying on a single deterministic continuation.

- Extensive experiments across four sports and surgical AQA benchmarks and 13 model–dataset combinations demonstrate consistent improvements over directly scoring truncated videos, with particularly substantial gains at early observation stages.

2 Related Work

Action quality assessment. Relevant AQA approaches study reference comparison, fine-grained execution modelling and probabilistic scoring. CoRe learns relative scores by comparing performances (Yu et al., 2021), while FineDiving and FineParser align action steps for fine-grained assessment (Xu et al., 2022b; 2024). USDL models score distributions to account for ambiguous annotations (Tang et al., 2020), and RICA^2 combines scoring rubrics with predictive uncertainty (Majeedi et al., 2024). Beyond individual action types, MoAKE studies unified assessment across actions (Xu et al., 2026a). These works improve how performances are represented and scored. Applying such models before a routine ends also requires reasoning about the unobserved execution.

Prediction from partial observations. Anytime AQA updates the predicted final score as more of a routine is observed; computational anytime classification instead varies the inference budget (Huang et al., 2018). Incomplete observations can also involve missing modalities, as studied by LIMSSR during training (Xu et al., 2026b). For partially observed videos, Activity Auto-Completion ranks training-video candidates for action recognition (Xu et al., 2015), while Vondrick et al. (2016) predict future visual representations. Online quality assessment includes detecting deviations from normal movement (Paiement et al., 2014) and providing immediate exercise feedback (Zhang et al., 2026). In sports scoring, Parmar & Morris (2017) initialise an LSTM with intermediate labels that increase towards the final score, then fine-tune it with final-score supervision. Gazis et al. (2026) also predict total surgical skill scores from truncated recordings. In this setting, estimating the final score requires accounting for how the unseen ending may affect performance quality.

Future modelling and conditional flows. One way to represent unseen execution is to predict future visual features. RIVER uses flow matching to generate future frames in a learned latent space, conditioned on past frames (Davtyan et al., 2023). Frozen Forecasting predicts future representations with diffusion and evaluates them through frozen downstream readouts (Walker et al., 2026). Continuous normalising flows also compute probability densities through invertible ODE dynamics (Chen et al., 2018), and flow matching trains their vector fields by regression (Lipman et al., 2023). Within AQA, flows have been used to improve representations: PHI reduces the gap between action-recognition and quality-assessment features (Zhou et al., 2025), while CaFlow combines bidirectional temporal refinement with counterfactual regularisation (Han et al., 2026). A^{2}QA extends a pretrained full-video scorer to early final-score prediction without retraining it. It combines scores from multiple completions formed with retrieved real endings, using weights derived from conditional probability densities.

3 Proposed Method

3.1 Problem Formulation and Preliminaries

Action quality assessment (AQA): given an activity video clip, the goal is to predict its performance score $y \in [0, s_{max}]$, where s_{max} is the maximum score.

Anytime AQA (A^{2}QA): given an activity video clip for scoring, let $x \in (0, 1]$ denote the observation ratio of the entire video length. A^{2}QA is to learn a scorer $\hat{y} : (0, 1] \rightarrow [0, s_{max}]$ which can score any partially observed video. For example, given a partial video with observation ratio x, we will learn to predict $\hat{y}(x)$, whose ground truth y is the score of full video. When the video is fully observed, i.e. $x = 1$, it becomes conventional AQA, with $\hat{y}(1) = S(F_{1:T})$, where S is the pretrained full-video scorer.

Following the conventional AQA framework (Zhou et al., 2026), the prediction pipeline contains three stages: a backbone extracts video features, a representation learning layer aggregates or transforms them, and a score prediction head outputs the score.

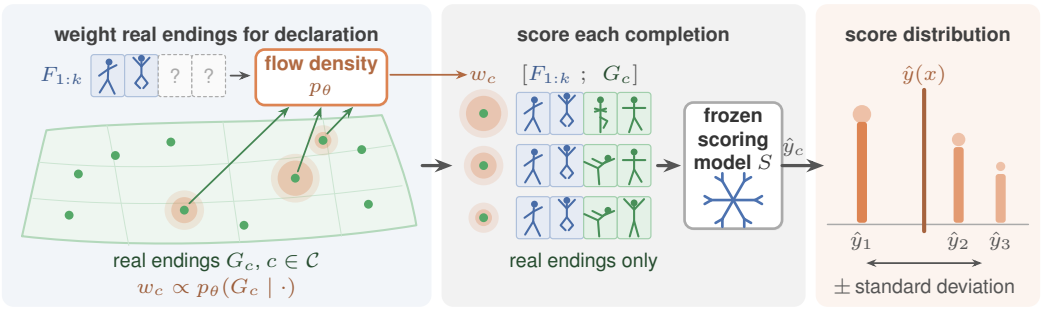

Figure 2: **Method overview.** Real endings are retrieved using the partial video and declaration. The displayed density weighting (Eq. 4) is refined into likelihood-ratio weights (Eq. 9). A frozen scorer evaluates the completed sequences; their scores and weights give the prediction and standard deviation. In $p_\theta(G_c \mid \cdot)$, the dot denotes conditioning on $(F_{1:k}, x)$.

3.2 Completing the partial video with retrieved real endings

To score a partial video x, we choose to first forecast its incomplete ending, and then score the whole video. Let T denote the number of non-overlapped short video segments that compose the video clip, and D the feature dimension. The backbone encodes the activity into feature embedding $F_{1:T} = [f_1; \ldots; f_T] \in \mathbb{R}^{T \times D}$, where each $f_i \in \mathbb{R}^D$ is the segment-level feature vector extracted from a fixed window of frames, and $[;]$ denotes the concatenation operation. Let $k = \lceil xT \rceil$; when $k = T$, we return $S(F_{1:T})$ directly.

Let $\mathcal{C}$ be a set of training videos that share the same category and declaration of the partial video x. For each video $c \in \mathcal{C}$, let $F^c \in \mathbb{R}^{T_c \times D}$ denotes the feature sequence, with T_c segments. Cutting it at the same ratio gives $k_c = \lceil xT_c \rceil$ and the corresponding ending $G_c = F^c_{k_c+1:T_c} \in \mathbb{R}^{(T_c - k_c) \times D}$. For $k < T$, we require each candidate ending to be nonempty. The completed video $[F_{1:k}; G_c]$ concatenates two partial videos along time, one from the input partial video and the other from the retrieved video ending. The *completion score* is $\hat{y}_c = S([F_{1:k}; G_c])$, where S is a pretrained scorer that can score a fully observed video.

Considering different video candidate $c \in \mathcal{C}$ may contribute differently to the prediction, we further learn the weight $w_c \in [0, 1]$ as *completion weight* of each candidate c, with $\sum_{c \in \mathcal{C}} w_c = 1$. For example, multiple retrieved endings may be coupled with the same observed beginning, we thus require to weight each real candidate c. We average the completion scores using these weights:

$$\hat{y}(x) = \sum_{c \in \mathcal{C}} w_c \, S([F_{1:k}; G_c]). \tag{1}$$

Following Bayesian rule, we can learn these weights w_c. For a video candidate c, its likelihood $\ell(G_c)$ measures how well candidate c explains the partial video x:

$$\ell(G_c) = p(F_{1:k} \mid G_c, x) \propto \ w_c. \tag{2}$$

Here $p(G_c \mid \cdot)$ denotes the probability assigned to candidate c, and $p(G_c \mid x)$ is its prior. With equal candidate priors, Bayes' rule gives

$$p(G_c \mid F_{1:k}, x) = \frac{\ell(G_c)\, p(G_c \mid x)}{p(F_{1:k} \mid x)} \propto \ell(G_c). \tag{3}$$

As a result, weights w_c should be also proportional to the posterior $p(G_c \mid F_{1:k}, x)$ with the uniform prior $p(G_c \mid x) = 1/|\mathcal{C}|$. In order to calculate the posterior, we leverage flow matching to learn conditional density $p_\theta(G_c \mid F_{1:k}, x)$ and estimate w_c:

$$w_c = \frac{p_\theta(G_c \mid F_{1:k}, x)}{\sum_{c' \in \mathcal{C}} p_\theta(G_{c'} \mid F_{1:k}, x)}. \tag{4}$$

3.3 Learning the completion weights w_c

The conditional density $p_\theta(G_c \mid F_{1:k}, x)$ in Eq. 4 is estimated in the flow matching model's feature space. The flow matching model learns from training videos, each split into a partial beginning and

its own ending. The scoring model S then continues to evaluate completed videos built from real endings. For a query video feature F, partially observed until $k < T$, the observed part is $F_{1:k} \in \mathbb{R}^{k\times D}$ and its own ending is $G = F_{k+1:T} \in \mathbb{R}^{(T-k)\times D}$. Each candidate video also contributes a real ending G_c for the density estimation of $p_\theta(G_c \mid F_{1:k}, x)$.

We train the flow model by conditional flow matching (Lipman et al., 2023). Draw a noise tensor $\epsilon \in \mathbb{R}^{(T-k)\times D}$ with independent standard Gaussian entries and a flow time $t \sim \mathrm{Uniform}[0,1]$. The interpolated state $G_t = (1-t)\epsilon + tG \in \mathbb{R}^{(T-k)\times D}$ lies on the straight path from noise to the true ending G. Let θ be a trainable real-valued parameter vector. The neural velocity field $v_\theta(G_t, t, F_{1:k}, x) \in \mathbb{R}^{(T-k)\times D}$ predicts the direction of motion conditioned on the partial video $F_{1:k}$, and observation ratio x. Flow matching learning objective match it to the path's velocity $G - \epsilon$:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\Big[\|v_\theta(G_t, t, F_{1:k}, x) - (G-\epsilon)\|_F^2\Big], \tag{5}$$

where $\|\cdot\|_F$ indicates the Frobenius norm. The expectation is over sampled observation ratios x of all training videos, together with noise and flow time. The learned field further defines the flow:

$$\frac{\mathrm{d}G_t}{\mathrm{d}t} = v_\theta(G_t, t, F_{1:k}, x), \qquad G_0 = \epsilon. \tag{6}$$

For a differentiable, invertible flow, let $p_\theta(G \mid F_{1:k}, x) : \mathbb{R}^{(T-k)\times D} \to (0,\infty)$ denote the conditional density of $G_1 = G$. For each candidate c, we integrate Eq. 6 backward from $G_1^{(c)} = G_c$ to recover the noise original of G_c:

$$G_0^{(c)} = G_c + \int_1^0 v_\theta(G_t^{(c)}, t, F_{1:k}, x)\,\mathrm{d}t. \tag{7}$$

Along this trajectory, $\frac{\mathrm{d}}{\mathrm{d}t}\log p_t(G_t^{(c)} \mid F_{1:k}, x) = -\operatorname{div}_G v_\theta(G_t^{(c)}, t, F_{1:k}, x)$ (Chen et al., 2018), where p_t is the density at time t. Integrating this relation from $t=1$ to $t=0$ and using the Gaussian density at $G_0^{(c)}$ gives

$$\log p_\theta(G_c \mid F_{1:k}, x) = \log p_0(G_0^{(c)}) + \int_1^0 \operatorname{div}_G v_\theta(G_t^{(c)}, t, F_{1:k}, x)\,\mathrm{d}t. \tag{8}$$

Here p_0 is the standard Gaussian density, and div_G denotes divergence with respect to the ending state. By integrating Eq. 8 into Eq. 4 we can finally obtain w_c.

Due to limited training data, we notice for some specific ending G, it may always bring high posterior regardless the partial observation of the beginning. Meanwhile, posterior $p_\theta(G_c \mid F_{1:k}, x)$ can be extremely high given specific partial video observation. In order to bring robust density estimation, we revise Eq. 4 to avoid such cases. Let $b_c = |\mathcal{C}|^{-1}\sum_{j\in\mathcal{C}} \log p_\theta(G_c \mid F'_j, x)$ be the mean log-density over reference training partial video F'_j at the same x. We subtract this reference value to obtain the rectified score $s_c = \log p_\theta(G_c \mid F_{1:k}, x) - b_c$. The rectified score compares each ending's log-density under the observed partial video with its mean log-density under training partial video in $\mathcal{C}$, providing a measure of relative compatibility.

On the other hand, in order to avoid extremely high posterior density, we apply a temperature $0 < \tau < \infty$ to smooth the completion weights:

$$w_c = \frac{\exp(s_c/\tau)}{\sum_{c'\in\mathcal{C}}\exp(s_{c'}/\tau)} = \frac{\exp((\log p_\theta(G_c \mid F_{1:k}, x) - b_c)/\tau)}{\sum_{c'\in\mathcal{C}}\exp((\log p_\theta(G_{c'} \mid F_{1:k}, x) - b_{c'})/\tau)}. \tag{9}$$

For finite τ, candidates with higher rectified scores receive greater weight, and τ controls how concentrated the weights are. If $\tau = \infty$, it becomes uniform weights $w_c = 1/|\mathcal{C}|$. We select the temperature from a fixed grid $\mathcal{T}$ by maximising Spearman correlation ρ_{train} on training data: $\tau_x^\star = \arg\max_{\tau\in\mathcal{T}} \rho_{\mathrm{train}}(\hat{\mathbf{y}}(x;\tau), \mathbf{y})$. Here $\hat{\mathbf{y}}(x;\tau)$ and $\mathbf{y}$ collect the predicted and ground-truth final scores of the training videos, respectively. The selected temperature is fixed for testing. The flow matching model is trained without score labels, and the scoring model remains frozen throughout.

The weighted standard deviation $\sigma(x) \in [0,\infty)$ of the completion scores measures their disagreement (Figure 3), which usually decreases with higher observation ratio x.

$$\sigma(x) = \left[\sum_{c\in\mathcal{C}} w_c\big(\hat{y}_c - \hat{y}(x)\big)^2\right]^{1/2}. \tag{10}$$

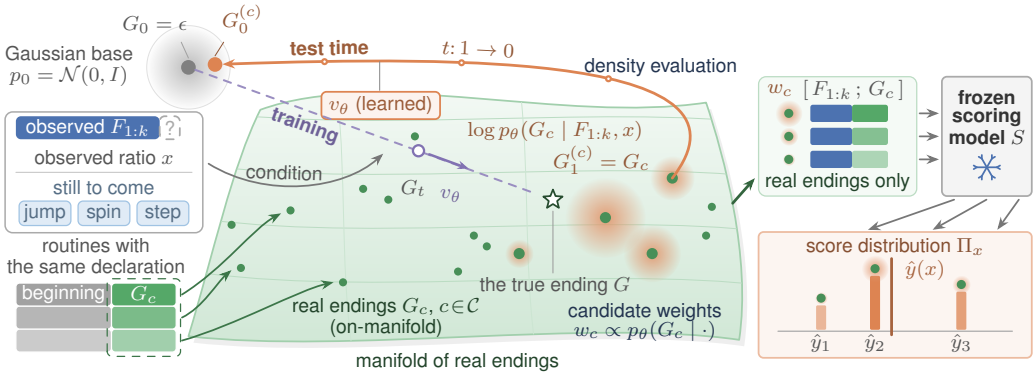

Figure 3: **Flow training and density evaluation.** Purple paths connect Gaussian noise to true endings during training (Eq. 5). Orange paths integrate backward from real candidates to recover $G_0^{(c)}$ and evaluate their conditional densities. Grey training noise and the orange recovered state lie in the same Gaussian base space. The proportionality shows Eq. 4; halos indicate the final weights (Eq. 9).

Meanwhile, a smaller $\sigma(x)$ also indicates a higher confidence of the scoring. For a chosen threshold $b \geq 0$, we may withhold the prediction when $\sigma(x) > b$, since it may not be reliable.

4 Experiments

4.1 Setup

Benchmarks and baselines. We evaluate eleven frozen scoring models in thirteen model–dataset pairs across four benchmarks. On FineFS (Ji et al., 2023) (figure skating), we use TSMP (Wang et al., 2025b), GDLT (Xu et al., 2022a) and B2S (Wang et al., 2025a). On FineDiving (Xu et al., 2022b) (diving), we use TSA (Xu et al., 2022b), FineCausal (Han et al., 2025), HP-MCoRe (Qi et al., 2025), CoRe (Yu et al., 2021), RICA^2 (Majeedi et al., 2024) and Uni-FineParser (Xu et al., 2025). We also evaluate CoRe and RICA^2 on JIGSAWS (Gao et al., 2014) (surgical skill), and MoAKE (Xu et al., 2026a) and LIMSSR (Xu et al., 2026b) on LapEx (Derathé et al., 2025) (laparoscopic exposure quality). We compare A^{2}QA (Ours) with directly scoring the partial video (Trunc.) using the same frozen scoring model. A^{2}QA completes the partial video with retrieved training endings and takes a weighted average of the completion scores. The model's prediction from the complete video (Full) serves as a reference. Table 1 reports predictions at $10/30/50/70/90\%$ without temporal averaging; Tables 2 and 3 report mean ρ over the same ratios.

Metrics. Following previous work (Yu et al., 2021; Xu et al., 2022b), we report Spearman's rank correlation $\rho = \frac{\sum_{i=1}^{N}(r_i-\bar{r})(\hat{r}_i-\bar{\hat{r}})}{\sqrt{\sum_{i=1}^{N}(r_i-\bar{r})^2 \sum_{i=1}^{N}(\hat{r}_i-\bar{\hat{r}})^2}}$ on all benchmarks. Here r_i and $\hat{r}_i$ are the ranks of the ground-truth and predicted scores of the N test routines, using average ranks for ties; $\bar{r}$ and $\bar{\hat{r}}$ are their means. On FineDiving, we also report the relative squared error (Yu et al., 2021), R-ℓ_2 $= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|y_i-\hat{y}_i|}{y_{\max}-y_{\min}}\right)^2$, where y_i and $\hat{y}_i$ are the ground-truth and predicted scores, and $y_{\max} - y_{\min}$ is the ground-truth score range. Spearman's ρ measures agreement in score rankings, whereas R-ℓ_2 measures normalised squared score error. We report R-$\ell_2 \times 100$; higher ρ and lower R-ℓ_2 indicate better performance.

Implementation. Each candidate pool contains at most 30 real training endings. Unless stated otherwise, we evaluate probability densities by integrating backward with 16 fourth-order Runge–Kutta (RK4) steps and estimate the divergence with 8 Rademacher probes. Reference correction and temperature fitting use only training data. For each video used to fit the temperature, candidate endings come from other training videos. The reference value b_c is also computed using prefixes from other training videos. Temperatures are fixed before testing.

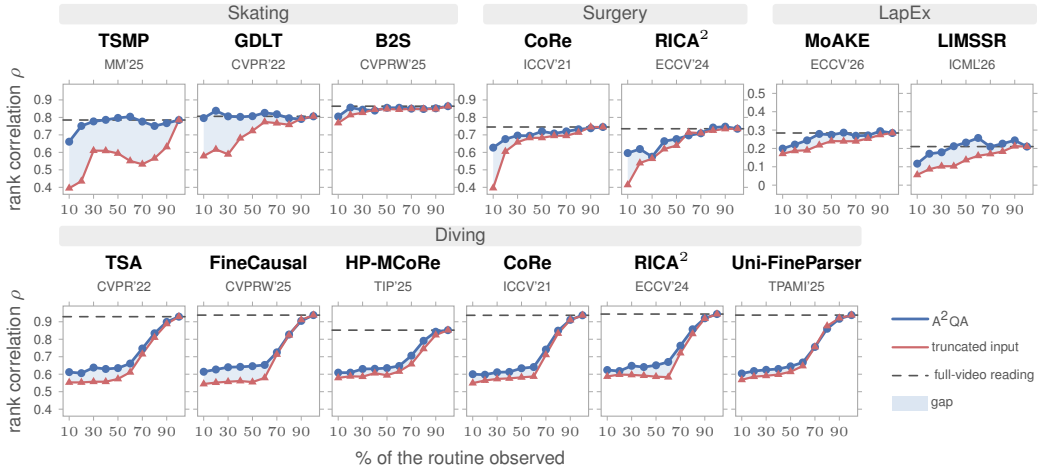

Figure 4: **Anytime scoring across thirteen model–dataset pairs.** Blue: A^{2}QA; red: direct partial-video scoring; shading: their gap; dashed: full-video scoring.

4.2 Anytime scoring across four domains

A^{2}QA improves ρ over direct partial-video scoring for all thirteen model–dataset pairs at 10%, 30% and 50% observation (Table 1 and Figure 4). For example, GDLT reaches its full-video correlation of 0.806 at 30%, compared with 0.589 for direct partial-video scoring. Across all 65 comparisons, 59 improve, one ties and five decrease at the displayed precision. The five decreases occur only at 70% or 90% and are no larger than 0.007. On FineDiving, R-ℓ_2 improves or ties in all comparisons except TSA at 90%. Figure 5 illustrates retrieved endings, while Figure 6 shows how the standard deviation of completion scores changes with the observation ratio.

The benefit varies across scorers within the same domain (Table 1). At 10% on FineFS, TSMP improves from 0.395 to 0.661 in ρ, while B2S improves from 0.768 to 0.805. B2S retains the higher correlation after completion despite the smaller gain. Improvements also persist at later observations for some models: at 90%, TSMP rises from 0.631 to 0.764, whereas B2S changes only from 0.848 to 0.852.

On FineDiving, Table 1 shows improvements in both ranking and numerical score accuracy. At 10%, TSA increases ρ from 0.553 to 0.612, while R-$\ell_2 \times 100$ falls from 4.73 to 1.30. At 50%, CoRe improves from 0.582 to 0.634 in ρ and reduces R-$\ell_2 \times 100$ from 1.81 to 1.19. The two metrics need not change together: at 90%, RICA^2 increases ρ from 0.918 to 0.920, while its error remains 0.30 at the reported precision. Completion can therefore improve both score ordering and score accuracy, with the size of each benefit depending on the model and observation ratio.

4.3 Flow matching and diffusion

We compare flow matching with diffusion for weighting the same retrieved endings (Table 2). Diffusion uses v-, ϵ- or x_0-prediction (Ho et al., 2020; Salimans & Ho, 2022). Both flow matching and diffusion use 1, 10 or 100 Euler steps, or 16 RK4 steps. All configurations use 8 Rademacher probes to estimate the divergence. Reference probability densities use the same model and integration settings as candidate evaluations; temperatures are fitted on training data. Diffusion probability densities are obtained from probability-flow ODEs (Song et al., 2021). The flow-matching RK4 column differs from Table 1's averages because Table 2 uses flow seed 0 instead of a three-seed mean on FineDiving and a subset of 466 training prefixes for reference correction and temperature fitting on FineFS.

Flow matching achieves the highest ρ on all six diving models and both JIGSAWS models, including ties. The best diffusion configurations achieve higher ρ on LapEx, while differences on skating are small. More integration steps do not consistently improve scoring. Given its strong overall performance, we use flow matching in A^{2}QA.

Table 1: **Early scoring on thirteen model–dataset pairs.** Trunc./Full: truncated/complete inputs to the same frozen model. Ours uses Eq. 9 weights. Bold: better paired value.

		10%		30%		50%		70%		90%		100%
Domain	Scoring model	Trunc.	Ours	Trunc.	Ours	Trunc.	Ours	Trunc.	Ours	Trunc.	Ours	Full
Spearman's ρ (↑)												
Skating	TSMP (Wang et al., 2025b)	0.395	**0.661**	0.610	**0.777**	0.594	**0.795**	0.532	**0.776**	0.631	**0.764**	0.785
	GDLT (Xu et al., 2022a)	0.579	**0.796**	0.589	**0.806**	0.723	**0.805**	0.767	**0.816**	**0.793**	0.791	0.806
	B2S (Wang et al., 2025a)	0.768	**0.805**	0.827	**0.842**	0.848	**0.857**	0.849	**0.851**	0.848	**0.852**	0.864
Diving	TSA (Xu et al., 2022b)	0.553	**0.612**	0.557	**0.638**	0.573	**0.635**	0.715	**0.747**	0.888	**0.899**	0.929
	FineCausal (Han et al., 2025)	0.544	**0.614**	0.557	**0.640**	0.555	**0.646**	0.713	**0.725**	**0.911**	0.905	0.938
	HP-MCoRe (Qi et al., 2025)	0.579	**0.609**	0.586	**0.630**	0.594	**0.635**	0.658	**0.706**	0.824	**0.843**	0.852
	CoRe (Yu et al., 2021)	0.549	**0.600**	0.573	**0.611**	0.582	**0.634**	0.711	**0.741**	0.910	**0.911**	0.937
	RICA^2 (Majeedi et al., 2024)	0.587	**0.624**	0.596	**0.648**	0.586	**0.651**	0.721	**0.763**	0.918	**0.920**	0.944
	Uni-FineParser (Xu et al., 2025)	0.568	**0.604**	0.591	**0.625**	0.614	**0.645**	0.756	0.756	**0.921**	0.917	0.938
Surgery	CoRe (Yu et al., 2021)	0.396	**0.627**	0.658	**0.696**	0.683	**0.720**	0.695	**0.720**	**0.745**	0.738	0.745
	RICA^2 (Majeedi et al., 2024)	0.414	**0.596**	0.564	**0.575**	0.639	**0.676**	**0.710**	0.709	0.734	**0.747**	0.735
LapEx	MoAKE (Xu et al., 2026a)	0.171	**0.199**	0.190	**0.244**	0.240	**0.274**	0.239	**0.269**	0.274	**0.294**	0.284
	LIMSSR (Xu et al., 2026b)	0.056	**0.117**	0.103	**0.179**	0.137	**0.232**	0.170	**0.209**	0.212	**0.244**	0.210
R-$\ell_2 \times 100$ (↓)												
Diving	TSA (Xu et al., 2022b)	4.73	**1.30**	4.49	**1.27**	4.11	**1.26**	2.11	**0.95**	**0.44**	0.45	0.357
	FineCausal (Han et al., 2025)	2.13	**1.23**	2.00	**1.21**	2.05	**1.18**	1.72	**0.92**	0.35	0.35	0.251
	HP-MCoRe (Qi et al., 2025)	3.03	**1.30**	2.55	**1.25**	1.96	**1.22**	1.51	**1.03**	0.74	**0.64**	0.600
	CoRe (Yu et al., 2021)	1.85	**1.26**	1.77	**1.23**	1.81	**1.19**	1.35	**0.90**	0.36	**0.34**	0.265
	RICA^2 (Majeedi et al., 2024)	1.53	**1.21**	1.57	**1.16**	1.65	**1.15**	1.17	**0.82**	0.30	0.30	0.222
	Uni-FineParser (Xu et al., 2025)	1.95	**1.31**	1.86	**1.26**	1.80	**1.21**	1.05	**0.87**	0.34	**0.33**	0.274

Table 2: **Flow matching and diffusion.** Mean ρ at $10/30/50/70/90\%$ observation using Euler ($1/10/100$ steps) or fourth-order Runge–Kutta (RK4, 16 steps). Bold: best value per row.

		Flow matching				Diffusion, v				Diffusion, ϵ				Diffusion, x_0			
Domain	Scoring model	1	10	100	RK4	1	10	100	RK4	1	10	100	RK4	1	10	100	RK4
Skating	TSMP (Wang et al., 2025b)	0.754	0.755	0.755	0.756	**0.757**	0.754	0.753	0.753	0.756	0.752	0.755	0.755	0.754	0.754	0.754	0.754
	GDLT (Xu et al., 2022a)	0.796	0.803	0.803	**0.804**	0.797	0.798	0.802	0.803	0.798	0.798	0.796	0.795	0.798	0.803	0.798	0.797
	B2S (Wang et al., 2025a)	0.835	0.840	**0.841**	**0.841**	0.836	0.835	0.840	0.840	0.836	0.836	0.835	0.835	0.837	0.839	0.838	0.837
Diving	TSA (Xu et al., 2022b)	0.698	0.702	0.706	**0.708**	0.696	0.697	0.700	0.703	0.696	0.696	0.697	0.697	0.698	0.698	0.697	0.697
	FineCausal (Han et al., 2025)	0.704	0.707	0.708	**0.709**	0.703	0.703	0.705	0.707	0.703	0.703	0.703	0.703	0.704	0.705	0.704	0.705
	HP-MCoRe (Qi et al., 2025)	0.680	0.681	**0.685**	**0.685**	0.679	0.679	0.680	0.681	0.679	0.679	0.679	0.679	0.680	0.680	0.681	0.681
	CoRe (Yu et al., 2021)	0.696	0.697	**0.699**	**0.699**	0.694	0.694	0.695	0.695	0.694	0.694	0.694	0.694	0.695	0.696	0.695	0.695
	RICA^2 (Majeedi et al., 2024)	0.715	0.718	0.721	**0.722**	0.714	0.715	0.717	0.717	0.714	0.714	0.715	0.715	0.715	0.716	0.716	0.715
	Uni-FineParser (Xu et al., 2025)	0.708	0.708	**0.710**	**0.710**	0.707	0.707	0.707	0.707	0.706	0.707	0.707	0.707	**0.710**	0.707	0.707	0.707
Surgery	CoRe (Yu et al., 2021)	0.699	0.700	**0.701**	0.700	0.695	0.694	0.696	0.696	0.663	0.660	0.661	0.661	0.644	0.641	0.638	0.640
	RICA^2 (Majeedi et al., 2024)	**0.669**	0.666	0.661	0.661	0.662	0.664	0.659	0.661	0.535	0.542	0.531	0.532	0.586	0.575	0.556	0.553
LapEx	MoAKE (Xu et al., 2026a)	0.276	0.257	0.256	0.256	0.278	0.265	0.254	0.253	**0.282**	0.274	0.274	0.274	0.274	0.274	0.278	0.278
	LIMSSR (Xu et al., 2026b)	0.236	0.195	0.195	0.196	0.257	0.213	0.188	0.186	**0.262**	0.248	0.250	0.249	0.240	0.232	0.234	0.234

Table 3: **Ablations.** Mean ρ over models and five ratios. Uniform: equal weights; Similarity: normalised positive prefix cosine similarities; Density: without subtracting b_c. (b) Ours fits τ per ratio, with $K = 8$ on diving and All on skating. Reference pools contain training prefixes that retrieve the ending (skating) or share its dive number (diving). K is the maximum number of reference training prefixes used to compute b_c for each ending. Subsets are nested in a fixed random order; All uses every available reference (up to 24 on diving). Bold: column best; shading: Ours.

(a) Score aggregation — Mean ρ (↑)

Setting	Skating	Diving
Uniform	0.698	0.669
Similarity	0.707	0.689
Density	**0.800**	0.690
Ours	**0.800**	**0.704**

(b) Parameters — Mean ρ (↑)

Setting	Skating	Diving
Ours	0.800	**0.704**
$\tau = 1$	0.744	0.590
$K = 1$	**0.801**	0.702
$K = 2$	0.800	0.703
$K = 4$	0.797	**0.704**
$K = 8$	0.797	**0.704**
All	0.800	**0.704**



Figure 5: **Endings selected by A^{2}QA.** Two examples per sport at 30%: FineDiving with RICA^2 (top) and FineFS with GDLT (bottom). Left: observed prefix. Right: the ending with highest conditional probability density p_θ above the true ending. GT: ground-truth score. Pred: the frozen model's completion score and the A^{2}QA prediction.

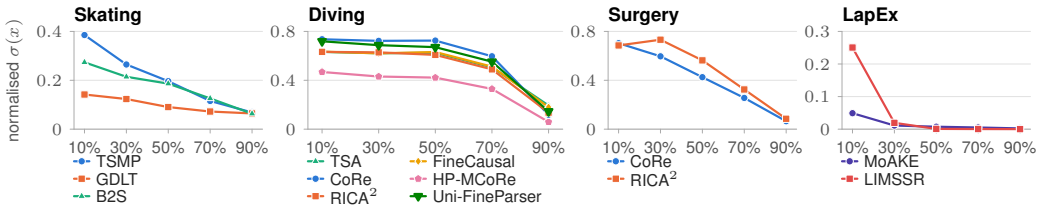

Figure 6: **Standard deviation of completion scores.** Each line averages a model's median test $\sigma(x)$ (Eq. 10) over seeds and divides by the ground-truth test-score standard deviation.

4.4 Ablations

Score aggregation. Table 3(a) favours Ours over Uniform and Similarity on both datasets. Subtracting b_c from the log density raises mean ρ on diving from 0.690 to 0.704; both skating settings round to 0.800. The gain from this correction is more apparent on diving.

Temperature and reference count. Fitting τ improves over $\tau = 1$ on skating (0.744 to 0.800) and diving (0.590 to 0.704; Table 3(b)). Across $K = 1, 2, 4, 8$ and All, mean ρ varies less: 0.797–0.801 on skating and 0.702–0.704 on diving. Temperature fitting thus has a larger effect here than the tested changes in reference count.

5 Conclusions

We presented A^{2}QA for anytime action quality assessment through partial video completion. Retrieved training endings represent possible continuations, and a conditional flow weights the frozen model's completion scores. This design separates future modelling from quality scoring. At full observation, A^{2}QA returns the original model's prediction. Across four benchmarks, A^{2}QA improves Spearman correlation over the same models scoring truncated videos for all thirteen model–dataset pairs at 10%, 30% and 50% observation. These results support partial video completion as a way to extend existing scoring models to earlier assessment without retraining.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and code organisation. We also consulted these tools for feedback on experimental comparisons and interpretation of results. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.07366.

Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *ICCV*, pp. 23263–23274, 2023. URL https://arxiv.org/abs/2211.14575.

Arthur Derathé, Fabian Reche, Sylvain Guy, Katia Charrière, Bertrand Trilling, Pierre Jannin, Alexandre Moreau-Gaudry, Bernard Gibaud, and Sandrine Voros. LapEx: A new multimodal dataset for context recognition and practice assessment in laparoscopic surgery. *Scientific Data*, 12:342, 2025. doi: 10.1038/s41597-025-04588-7.

Yixin Gao, S. Swaroop Vedula, Carol E. Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C. Lin, Lingling Tao, Luca Zappella, Benjamín Béjar, David D. Yuh, Chi Chiung Grace Chen, René Vidal, Sanjeev Khudanpur, and Gregory D. Hager. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, 2014. URL https://cirl.lcsr.jhu.edu/wp-content/uploads/2015/11/JIGSAWS.pdf.

Athanasios Gazis, Dimitrios Schizas, Stylianos Kykalos, Pantelis Karaiskos, and Constantinos Loukas. Egocentric video analysis for automated assessment of open surgical skills via deep learning. *International Journal of Computer Assisted Radiology and Surgery*, 21(2):297–306, 2026. doi: 10.1007/s11548-025-03518-7.

Ruisheng Han, Kanglei Zhou, Amir Atapour-Abarghouei, Xiaohui Liang, and Hubert P. H. Shum. FineCausal: A causal-based framework for interpretable fine-grained action quality assessment. In *CVPR Workshops*, pp. 6064–6073, 2025. doi: 10.1109/CVPRW67362.2025.00599.

Ruisheng Han, Kanglei Zhou, Shuang Chen, Amir Atapour-Abarghouei, and Hubert P. H. Shum. CaFlow: Enhancing long-term action quality assessment with causal counterfactual flow. In *WACV*, pp. 8231–8241, 2026. doi: 10.1109/WACV61042.2026.00794. URL https://arxiv.org/abs/2511.21653.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018. URL https://arxiv.org/abs/1703.09844.

Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted uncertainty score disentanglement network for action quality assessment. In *ACM MM*, pp. 8590–8597, 2023. doi: 10.1145/3581783.3613795.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. RICA^2: Rubric-informed, calibrated assessment of actions. In *ECCV*, volume 15121, pp. 143–161, 2024. doi: 10.1007/978-3-031-73036-8_9.

Adeline Paiement, Lili Tao, Sion Hannuna, Massimo Camplani, Dima Damen, and Majid Mirmehdi. Online quality assessment of human movement from skeleton data. In *BMVC*, 2014. doi: 10.5244/C.28.79. URL https://bmva-archive.org.uk/bmvc/2014/files/paper058.pdf.

Paritosh Parmar and Brendan Tran Morris. Learning to score Olympic events. In *CVPR Workshops*, pp. 20–28, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017_workshops/w2/html/Parmar_Learning_to_Score_CVPR_2017_paper.html.

Mengshi Qi, Hao Ye, Jiaxuan Peng, and Huadong Ma. Action quality assessment via hierarchical pose-guided multi-stage contrastive regression. *IEEE Transactions on Image Processing*, 34:6461–6474, 2025. doi: 10.1109/TIP.2025.3613952. URL https://arxiv.org/abs/2501.03674.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

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 *CVPR*, pp. 9839–9848, 2020. doi: 10.1109/cvpr42600.2020.00986.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, pp. 98–106, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Vondrick_Anticipating_Visual_Representations_CVPR_2016_paper.html.

Jacob C. Walker, Pedro Vélez, Luisa Polania Cabrera, Guangyao Zhou, Sayna Ebrahimi, Rishabh Kabra, Carl Doersch, Maks Ovsjanikov, João Carreira, and Shiry Ginosar. Frozen forecasting: A unified evaluation. arXiv preprint arXiv:2507.13942v2, 2026. URL https://arxiv.org/abs/2507.13942v2. Version 2, 15 April 2026.

Fengshun Wang, Qiurui Wang, and Dan Chen. From beats to scores: A multi-modal framework for comprehensive figure skating assessment. In *CVPR Workshops*, pp. 5951–5960, 2025a. doi: 10.1109/cvprw67362.2025.00588.

Fengshun Wang, Qiurui Wang, and Peilin Zhao. Learning long-range action representation by two-stream Mamba pyramid network for figure skating assessment. In *ACM International Conference on Multimedia*, pp. 867–875, 2025b. doi: 10.1145/3746027.3754702.

Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *CVPR*, pp. 3232–3241, 2022a. doi: 10.1109/cvpr52688.2022.00323.

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 *ECCV*, volume 17005, pp. 294–312, 2026a. doi: 10.1007/978-3-032-37369-4_16. URL https://link.springer.com/chapter/10.1007/978-3-032-37369-4_16. arXiv:2607.19826.

Huangbiao Xu, Huanqi Wu, Xiao Ke, and Yuxin Peng. LIMSSR: LLM-driven sequence-to-score reasoning under training-time incomplete multimodal observations. In *ICML*, 2026b. URL https://icml.cc/virtual/2026/poster/66773. arXiv:2605.00434.

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 *CVPR*, pp. 2949–2958, 2022b. doi: 10.1109/cvpr52688.2022.00296.

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 *CVPR*, pp. 14628–14637, 2024. doi: 10.1109/cvpr52733.2024.01386.

Jinglin Xu, Sibo Yin, and Yuxin Peng. Human-centric fine-grained action quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6242–6255, 2025. doi: 10.1109/tpami.2025.3556935.

Zhen Xu, Laiyun Qing, and Jun Miao. Activity auto-completion: Predicting human activities from partial videos. In *ICCV*, pp. 3191–3199, 2015. doi: 10.1109/ICCV.2015.365.

Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *ICCV*, pp. 7919–7928, 2021. doi: 10.1109/iccv48922.2021.00782.

Yongchang Zhang, Boxuan Xu, Zhaowen Lin, Junjie Li, and Anlong Ming. Exercise quality assessment in monocular video streaming. *Engineering Applications of Artificial Intelligence*, 177:114905, 2026. doi: 10.1016/j.engappai.2026.114905.

Kanglei Zhou, Hubert P. H. Shum, Frederick W. B. 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. doi: 10.1109/TIP.2025.3574938.

Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert P. H. Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *Pattern Recognition*, 179:113933, 2026. doi: 10.1016/j.patcog.2026.113933. URL https://arxiv.org/abs/2412.11149v2.

A MORE QUALITATIVE RESULTS

Figures 7–9 show ten test examples each from FineFS, FineDiving and JIGSAWS at 30% observation. Each example shows the partial video on the left. On the right, the retrieved ending with the highest weight appears above the true ending. GT denotes the ground-truth score. Pred denotes the frozen model's completion score in the top row and the A^2QA prediction in the bottom row.



Figure 7: **Skating (FineFS, B2S).** Columns show the first four declared elements still to come. Each retrieved element is shown above the true element of the same class.

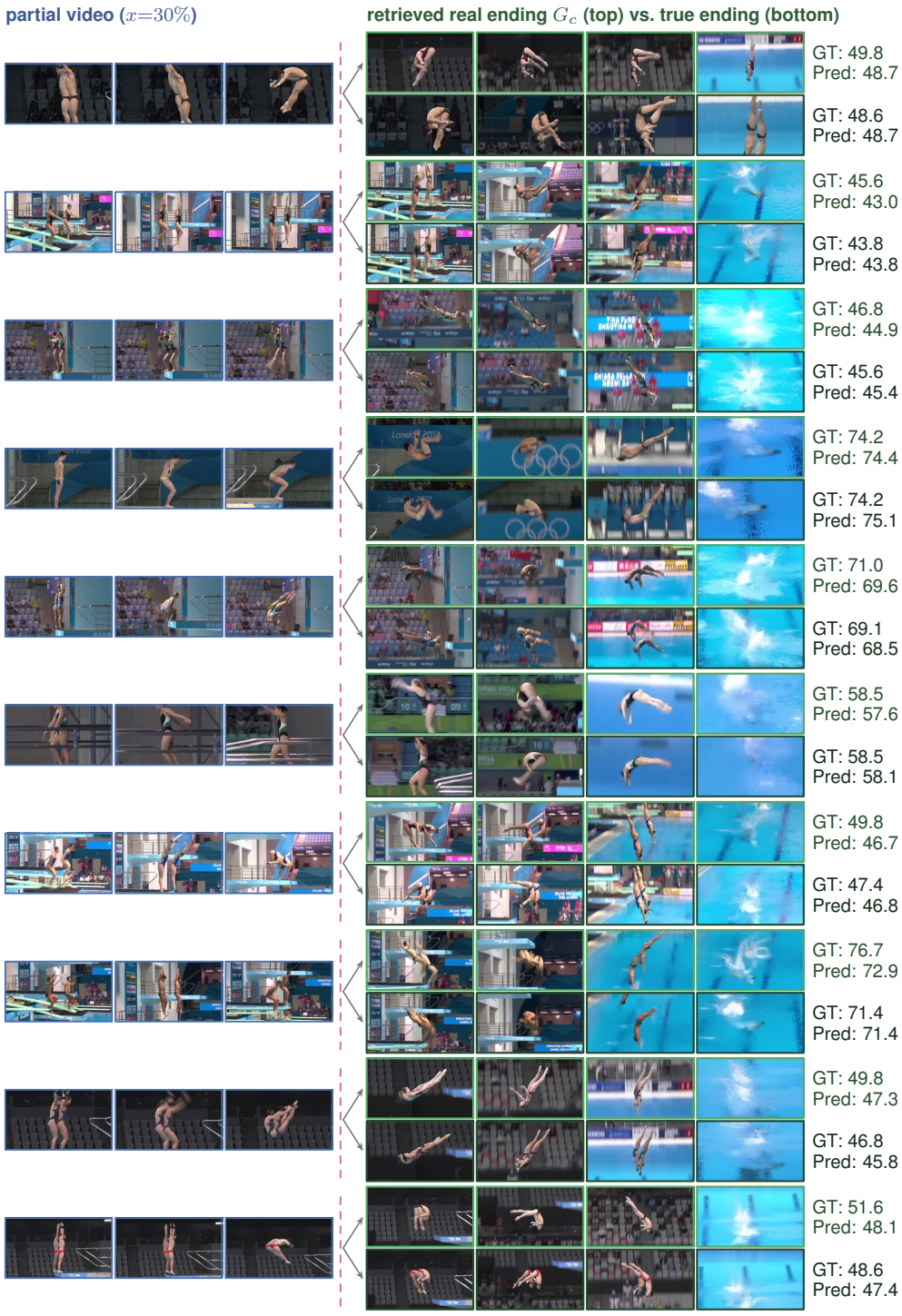

Figure 8: **Diving (FineDiving, TSA).**

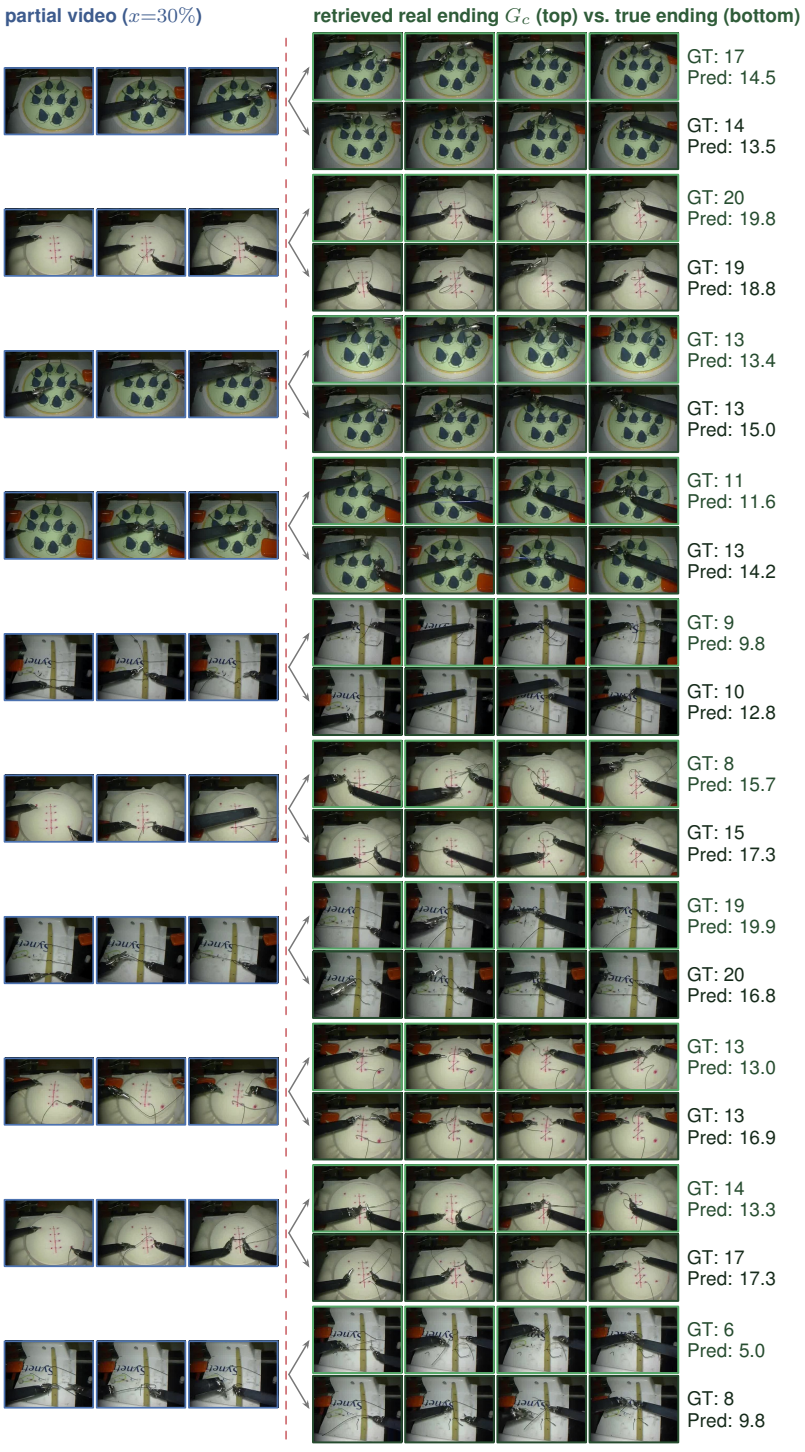

Figure 9: **Surgery (JIGSAWS, CoRe).** GT denotes the global rating score.