ProSkill: Segment-Level Skill Assessment in Procedural Videos

Michele Mazzamuto*
University of Catania, Italy

Daniele Di Mauro*
Next Vision s.r.l., Italy

Gianpiero Francesca†
Toyota Motor Europe, Belgium

Giovanni Maria Farinella†
University of Catania, Italy

Antonino Furnari†
University of Catania, Italy

Abstract

Skill assessment in procedural videos is crucial for the objective evaluation of human performance in settings such as manufacturing and procedural daily tasks. Current research on skill assessment has predominantly focused on sports and lacks large-scale datasets for complex procedural activities. Existing studies typically involve only a limited number of actions, focus on either pairwise assessments (e.g., A is better than B) or on binary labels (e.g., good execution vs needs improvement). In response to these shortcomings, we introduce ProSkill*, the first benchmark dataset for action-level skill assessment in procedural tasks.* ProSkill *provides absolute skill assessment annotations, along with pairwise ones. This is enabled by a novel and scalable annotation protocol that allows for the creation of an absolute skill assessment ranking starting from pairwise assessments. This protocol leverages a Swiss Tournament scheme for efficient pairwise comparisons, which are then aggregated into consistent, continuous global scores using an ELO-based rating system. We use our dataset to benchmark the main state-of-the-art skill assessment algorithms, including both ranking-based and pairwise paradigms. The suboptimal results achieved by the current state-of-the-art highlight the challenges and thus the value of* ProSkill *in the context of skill assessment for procedural videos. All data and code are available at* https://fpv-iplab.github.io/ProSkill/.

1. Introduction

Skill assessment is a fundamental task in human activity understanding, as it enables the objective evaluation of how well a person performs a given task. This capability is particularly valuable in procedural settings such as manufacturing and assembly, where the assessment of employee skills can directly impact efficiency, safety, and quality. In addition, wearable devices offer a promising avenue for skill assessment, enabling continuous, real-time monitoring of performance for early detection of errors or suboptimal behavior [9, 19, 29, 30]. Previous research on human skill assessment has been predominantly focused on sports [21,26,33,37,38], where datasets and evaluation protocols are more mature and readily available. Some efforts have extended skill assessment to other contexts, including surgical procedures [6,16,17] and egocentric activities [13]. Unfortunately, these datasets are generally limited in scale, task diversity, and procedural structure. A recent step forward is represented by Ego-Exo4D [12], which captures

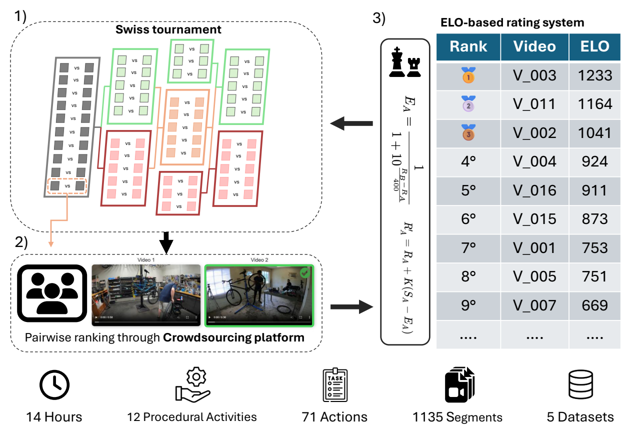


Figure 1. The proposed annotation protocol (top) and key features of the resulting ProSkill dataset (bottom). **Stage 1** action video pairs are selected for labeling following a Round of a Swiss Tournament scheme; **Stage 2** selected pairs are labeled with a crowdsourcing platform asking users to perform pairwise ranking; **Stage 3** the pairwise outcomes are aggregated using an ELO [7] scheme to compute a global leaderboard. Stages 1-3 are iterated for a given number of rounds to achieve stable absolute ratings.

*Equal contribution.
†Co-Principal Investigator role.

goal-driven activities from both egocentric and exocentric perspectives, providing segment-level binary skill labels for a subset of videos. Despite these efforts, a large-scale, annotated dataset offering skill-level annotations across diverse real-world tasks is currently missing.

To address this limitation, we introduce PROSKILL (Figure 1) the first dataset specifically designed for procedural skill assessment, with the aim of supporting the development and evaluation of models that estimate human expertise in structured tasks. Previous skill assessment datasets have typically adopted either absolute scores [11, 22, 25–27, 30, 37] or pairwise comparisons [5, 6, 13] to asses skill. Absolute scoring allows direct skill estimation and supports regression-based evaluation, but is often ambiguous and inconsistent outside of sports and other constrained domains where trained judges can apply objective criteria (e.g., body positions). Pairwise comparisons, on the other hand, are more intuitive and reliable for annotators, as they reduce ambiguity by asking to evaluate which of two performances is better. However, they do not inherently provide globally interpretable skill scores, which are essential for actionable feedback (e.g., "above average" or "needs improvement"). PROSKILL bridges this gap by combining pairwise comparisons into globally consistent ranking, allowing for multiple evaluation paradigms (ranking, regression, and classification), within a dataset.

To achieve this, we propose a novel annotation protocol consisting of three stages based on a Swiss Tournament scheme, crowdsourcing annotation and ELO rating system [7]. Swiss Tournament and ELO are long-established practices in competitive domains such as chess. This approach avoids the cost of exhaustive pairwise comparisons while preserving robustness. PROSKILL consists of recordings of diverse tasks performed in unconstrained, real-world conditions. Such recordings are gathered from publicly available video datasets that have been widely used for procedural activity understanding. Specifically, PROSKILL includes curated annotations on sequences from *EgoExo4D* [12], *Meccano* [28], *EpicTent* [14], *Ikea ASM* [3], and *Assembly101* [30].These datasets cover diverse tasks, environments, and viewpoints, making PROSKILL a comprehensive benchmark for skill assessment across multiple domains. Using PROSKILL, we evaluate several state-of-the-art algorithms, including both global- and pairwise-ranking models. Results indicate significant room for improvement, confirming that procedural skill assessment remains a challenging open problem.

In summary, the main contributions of this work are: 1) A novel annotation protocol based on a Swiss Tournament, combining crowd-sourced pairwise annotations and ELO rating for scalable collection of absolute skill scores; 2) PROSKILL, the first dataset of procedural activities spanning multiple domains and environments, with both absolute and relative skill annotations; 3) A benchmark of state-of-the-art methods for pairwise and global ranking on diverse real-world tasks. All labels, code for the annotation protocol, and experimental pipelines will be publicly released.

2. Related Works

Datasets for Skill Assessment Skill assessment from video has been explored across a variety of domains, including sports, medical training, rehabilitation, music, and procedural tasks. The resulting datasets differ significantly in their annotation strategies, ranging from absolute scores to pairwise comparisons, as well as in activity structure and data modality. This diversity reflects a fragmented landscape, where progress in one area often does not transfer easily to others. Initial efforts concentrated on competitive sports, where performance is inherently linked to objective scoring. Datasets such as MIT-Dive [27], MIT-Skate [27], UNLV-Dive [25], UNLV-Vault [25], AQA-7 [22], and MTL-AQA [26] associate skill with scores assigned during official competitions. More recent datasets such as FineDiving [38], LOGO [41], and FineFS [15] improve annotation consistency by leveraging expert consensus, multi-view video, and more diverse subjects and actions. Beyond sports, structured domains such as surgery and rehabilitation have received attention due to their reliance on expert-defined protocols and high-stakes outcomes. JIGSAWS [11] and KIMORE [15], for example, provide fine-grained demonstrations of surgical or rehabilitative tasks annotated using clinical frameworks like OSATS [18]. These datasets often include multimodal data (e.g., RGB-D, motion capture), supporting detailed analysis of motor control and execution. Moving further into everyday and creative domains, datasets such as EPIC-Skills [5], BEST [6], and Piano-Skills [24] sidestep the subjectivity and cost of absolute scoring by adopting pairwise comparisons. This approach simplifies annotation and has been shown to scale effectively while maintaining robustness in skill ranking.

More recently, focus has shifted toward procedural and assembly tasks, where activities are long, hierarchical, and goal-directed. These tasks demand a deeper understanding of temporal structure, decision points, and success criteria. Assembly101 [30] provides egocentric recordings with subject-level self-assessments. EPIC-Tent [14] captures egocentric tent assembly performances annotated with self-rated uncertainty scores. EgoExoLearn [13] adopts pairwise comparisons over video segments to capture relative proficiency. The Struggle Determination dataset [8] takes a different perspective by labeling segments where users appear to struggle, surfacing cues like hesitation or error that are indicative of lower skill. While these datasets mark progress toward real-world skill assessment in proce-

Name	Year	Domain	Skill Annotation			Samples	Hours	Actions
			Type	Source	Granularity			
Datasets with Absolute Score Annotations								
MIT-Dive [27]	2014	Sport	Absolute Score	Judges	Video	159	0.12	1$*$
MIT-Skate [27]	2014	Sport	Absolute Score	Judges	Video	150	7.29	1
JIGSAWS [11]	2014	Surgery	Absolute Score	Experts	Video	103	$\sim$10	3
UNLV-Dive [25]	2017	Sport	Absolute Score	Judges	Video	717	7.10	1$*$
UNLV-Vault [25]	2017	Sport	Absolute Score	Judges	Video	176	0.12	1
AQA-7 [22]	2019	Sport	Absolute Score	Judges	Video	1106	$\sim$2	7
MTL-AQA [26]	2019	Sport	Absolute Score	Judges	Video	1412	1.6	2$*$
Fis-V [37]	2019	Sport	Absolute Score	Judges	Video	500	22	1
Assembly101 [30]	2022	Assembly	Absolute Score	Self-reported	Subject	53	513	202
Datasets with Other Annotation Types (Uncertainty level, Difficulty, Categorical, Binary, etc.)								
EpicTent [14]	2019	Assembly	Uncertainty	Self-reported	Segment	34090	5.40	38
Piano-Skills [24]	2021	Music	Grade, Difficulty	Teachers	Video	992	$\sim$3	9$\star$
HoloAssist [36]	2023	Manipulation	Self-Reported Skill	Self-reported	Subject	222	169	414
Ego-Exo4D [12]	2024	Assembly + Daily	Categorical Labels Binary Labels	Surveys Experts	Subject Segment	740 2539	1286	NA
Datasets with Pairwise Ranking Annotations								
EPIC-Skill [5]	2018	Surgery + Daily	Pairwise Ranking	Crowdsourcing	Video	216	5.20	4
BEST [6]	2019	Daily	Pairwise Ranking	Crowdsourcing	Video	500	26	5
EgoExoLearn [13]	2024	Cooking	Pairwise Ranking	Crowdsourcing	Segment	3304	$\sim$9	4
PROSKILL	**2025**	**Procedural**	**Absolute + Pairwise**	**Crowdsourcing**	**Segment**	**1135**	**14**	**71**

Table 1. Datasets with skill annotations. **PROSKILL** provides fine-grained **segment-level** annotations on procedural videos with both **absolute and pairwise** skill ratings, covering 14 **hours** of video and 71 **diverse actions**. $*$: Dive height. $\star$: Song difficulty levels.

dural videos, they have several limitations: small scale (few subjects or short videos), limited scenarios (constrained environments or narrow tasks), coarse labels (binary or clip-level instead of continuous or subject-level), low action variability, and lack of scalable annotation protocols (relying on self-reports or expert judgments). These factors limit generalization across tasks and hinder large-scale extension.

Differently from prior efforts, PROSKILL focuses on procedural activities annotated at the segment level, enabling more fine-grained assessment of user performance. It provides both pairwise skill comparisons and globally consistent absolute scores, covering multiple real-world domains. As reported in Table 1, PROSKILL is diverse, including 1135 segments across 71 actions and 14 different hours of video, exceeding similar features of previous datasets related to sports, surgery, or music domains. While some benchmarks provide larger numbers of hours (e.g., Ego-Exo4D, Assembly101 and HoloAssist), or segments (e.g., EPIC-Tent and Ego-Exo4D), they either tackle the coarser-grained subject level, provide only binary labels, cover a very small number of actions, or report subjective labels (e.g., self-rated uncertainty) that limit their applicability. All in all, PROSKILL offers a unique combination of scale, diversity, and annotation depth, setting a new standard for the fine-grained analysis of procedural skills.

Global Ranking Methods Early Action Quality Assessment (AQA) approaches, such as Parmar et al. [23], leveraged C3D features with SVR or LSTM regressors, aiming to balance data efficiency with temporal modeling. To mitigate label ambiguity, USDL [31] proposed modeling scores as probability distributions rather than point estimates, extended by MUSDL to incorporate multi-path supervision. TSA-Net [34] improves contextual understanding through Tube Self-Attention focused on foreground motion, achieving efficient and competitive performance. DAE-AQA [40] adopts a variational autoencoder to learn score distributions and capture uncertainty, enhancing robustness to subtle quality differences. CoFInAl [42] formulates AQA as a hierarchical coarse-to-fine classification problem, integrating temporal fusion and grade parsing to support fine-grained, generalizable predictions. More recently, Okamoto et al. [20] introduced a neuro-symbolic framework that extracts interpretable action descriptors and applies symbolic reasoning, enabling transparent and expert-aligned evaluations.

In this paper, we benchmark the main representatives of these approaches on PROSKILL, showing that they achieve limited performance when confronted with real-world videos of procedural activities. This highlights that future research efforts should be dedicated to further developing absolute skill assessment approach in the domain of procedural video, which PROSKILL enables.

Pairwise Ranking Methods Skill assessment has also been approached as a pairwise ranking task, where the goal is to determine which of two videos demonstrates a higher level of skill. The work of [5] proposes a deep learning method to rank skill levels from videos using a novel loss function that adapts to skill differences, achieving 70–83% accuracy across tasks like surgery and drawing, enabling automated skill assessment and video organization. The RAAN model [6] assesses relative skill in long videos using a learnable temporal attention mechanisms. The model focuses on identifying and attending to the most skill-relevant segments to determine overall performance. RAAN is trained using video-level supervision and a rank-aware loss function. CoRe [39] predicts relative quality by contrasting a target video against a reference video with comparable features. The Contrastive Regression framework learns quality distinctions through paired video comparisons, while a group-aware regression tree is used to hierarchically regress the correct score. AQA-TPT [1] introduces a temporal parsing transformer that decomposes holistic video representations into temporally segmented part-level features, enabling the capture of fine-grained variations within the same action class. For quality scoring, the method leverages a contrastive regression approach applied to these part-based features.

We evaluate the performance of the main pairwise ranking approach on PROSKILL, showing their limitations, when tested on real-world procedural videos and highlighting the need for future investigations in this area.

3. The PROSKILL Annotation Protocol

We annotate PROSKILL iterating through three stages, as highlighted in Figure 1: stage 1) video pairs are selected for labeling following a Swiss Tournament scheme; stage 2) pairwise ranking labels are crowd-sourced for the selected pairs using Amazon Mechanical Turk (AMT); stage 3) the ELO ranking system is used to convert relative pairwise rankings into absolute scores, obtaining an absolute ranking of all labeled clips. Stages 1-3 are iterated for R rounds to obtain stable absolute rankings. PROSKILL, as a result, provides both relative and absolute skill annotations within a unified framework, enabling flexible evaluation protocols and supporting a broader range of training objectives. The proposed annotation approach is scalable to large datasets, robust to noisy annotations, and, requiring no expert involvement, it can be annotated through a crowd-sourced protocol, such as Amazon Mechanical Turk (AMT), enabling efficient and reliable skill estimation across diverse procedural domains. Below, we describe each stage of the annotation pipeline in detail.

Stage 1 This stage follows a round of a Swiss Tournament, a non-elimination tournament format commonly used in popular games such as Chess, where maintaining a large number of players throughout the tournament is desirable. In our settings video segments are the players. Video segments are paired in each round based on their current standings (e.g., absolute scores coming from a previous stage), with those having similar scores facing each other whenever possible. Unlike single-elimination tournaments, where a single loss eliminates a participant, in a Swiss Tournament every video segment competes in every round, typically lasting a predetermined number of rounds. The pairing algorithm ensures that video segment face opponents of similar skill levels as the tournament progresses, while avoiding matching two opponents multiple times in the same tournament. This format provides a fair and comprehensive assessment of video segment skill across multiple games while maximizing playing time for all participants. We treat a pairwise comparison between two segments of the same action as a match. At the first iteration, every segment start with an equal score, while scores are gradually updated in Stage 3. The result of this stage is a set of matched video segments, for which we should obtain pairwise ranking labels. Note that, by matching segments with similar absolute scores, we can schedule for labeling only the most informative segment pairs, avoiding trivial comparisons.

Stage 2 we used Amazon Mechanical Turk (AMT) to label the video pairs selected in stage 1. Each annotation indicated which of the two "opponents" exhibited higher skill, used to refine the absolute skill score of each clip.

We designed a generalizable annotation pipeline to efficiently collect reliable pairwise skill judgments at scale. Although implemented on AMT, the methodology is platform-agnostic and can be adapted to any system supporting video playback, qualification gating, and structured task delivery. To encourage reproducibility, we will release the code and templates used in our crowdsourcing setup.

To ensure quality, workers were pre-screened based on performance history. Only contributors with a historical approval rate above 90% and who passed a dedicated qualification test were allowed. The test consisted of gold-standard comparisons, manually curated and updated three times across stages per dataset, to evaluate the worker's ability to reliably discern skill differences.

Each annotation task included a set of video pairs. Workers watched both videos and indicated which performer demonstrated higher skill. Instructions included positive and negative examples. The interface was designed to focus attention on the comparison itself (see Fig. 2). Each pair was annotated by five independent workers, with the final outcome computed by majority vote. In cases of low agreement, the pair was re-assigned to three additional workers. The output of this stage is a set of pairwise labels for previously matched video segments. Each label can be interpreted as the outcome of the associated match, stating which of the two segment "wins" the match (i.e., has the highest skill score). These labels will be used in stage 3 to obtain a consistent global ranking among segments. Following the outlined procedure, overall, we collected 16,372

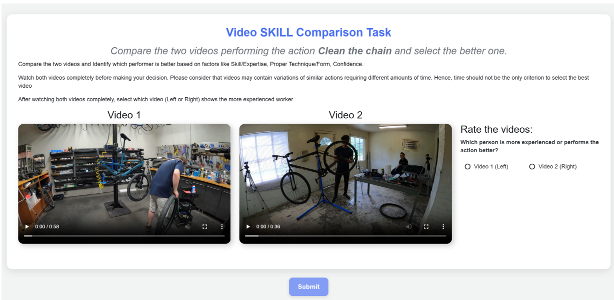


Figure 2. Screenshot of the annotation interface used in AMT. Annotators view two video segments corresponding to the same task and are asked to select which performer appears more skilled.

unique comparisons, annotated by 551 qualified workers, which have been used to produce the proposed PROSKILL benchmark Dataset. See the supplementary material for further details on the AMT platform implementation and agreement rate statistics.

Stage 3 The goal of this stage is to update the absolute ranking of video segments using the the ELO rating system [7]. The system was developed by physicist Arpad Elo to calculate the relative skill levels of players in zero-sum games such as Chess. In our case ELO assigns each video segment a numerical rating that increases with wins and decreases with losses. The magnitude of the rating's change is determined by the difference in ratings between opponents and the actual outcome versus the expected outcome. The expected score for a video segment is calculated using the logistic function:

$$E_A = \frac{1}{1+10^{(R_B - R_A)/400}} \quad E_B = \frac{1}{1+10^{(R_A - R_B)/400}} \tag{1}$$

where R_A and R_B are the current absolute ratings of video segment A and B respectively. After a match, the new rating is computed as

$$R'_A = R_A + K(S_A - E_A) \tag{2}$$

where K is a constant that determines the maximum possible rating change, and S_A is the actual score (1 for a win, 0.5 for a draw, 0 for a loss). This self-correcting system ensures that ratings accurately reflect current playing strength. In our setting, this method permits to assign a global expertise level to each clip, without the necessity for annotators to globally rank all the clips. In our settings, we start from video segment pairs (A, B) annotated in stage 2, where $K = 1$ if A is selected over B and $K = 0$ otherwise. We obtain R_A and R_B as the absolute scores of segments A and B from the previous round, initialized to zero before the first round. Absolute scores are hence updated using (1)-(2). Once the absolute scores are updated, we can return to stage 1 to select informative segment pairs to label for pairwise ranking.

4. The PROSKILL Dataset

Data Selection PROSKILL is based on established, public datasets of procedural videos, to which we add segment-level skill annotations following the proposed protocol. We chose these datasets because they provide clear examples of procedural activities where expertise is clearly assessable, and they complement each other. Specifically, we consider the following datasets:

- *EgoExo4D* [12]: we focused on bike repair data selecting 4 scenarios for a total of 191 clips across 12 actions.
- *Meccano* [28]: we sourced segment annotations from EgoProcel [2] annotations - the dataset has a single scenario for a total of 80 clips across 5 actions.
- *EpicTent* [14]: we sourced segment annotations from EgoProcel [2] annotations - the dataset has a single scenario for a total of 144 clips across 9 actions.
- *IkeaASM* [3]: we selected 4 scenarios for a total of 160 clips across 10 actions.
- *Assembly101* [30]: we selected 2 scenarios for a total of 560 clips across 35 actions.

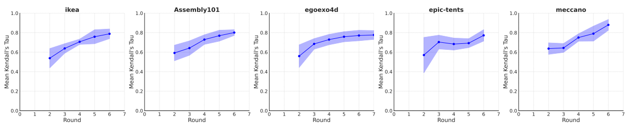
For each action, we select 16 clips, which are labeled for pairwise and absolute skill scores using the proposed annotation protocol. Note that, since comparing segments depicting different actions is unfeasible, we label segments from each action independently, obtaining a different absolute ranking for each action. Absolute scores are hence converted to percentiles to be mapped to the same scale.

Annotation Details We set the number of rounds $R = 6$. This number exceeds the minimum recommended number of rounds, which is usually set to the base 2 logarithm of the number of participants [10], which is set to 16 segment per action in our case. Following this scheme, we annotated approximately 40% of all possible video segment pairings for each action. While more rounds and labeled pairs generally lead to finer-grained differentiation, we observe that 6 rounds already provides quite stable results, while significantly reducing the number of comparisons. To quantitatively assess ranking quality, we use Kendall's tau, a rank correlation coefficient that measures the ordinal association between two variables and indicates the similarity of the orderings (rankings) of the two sets. Figure 3a reports Kendall's τ coefficients computed between rankings at consecutive rounds, which measures ranking consistency as annotation progresses. Datasets such as *IKEA*, *Assembly101* and *EgoExo4D*, exhibit a clear flattening trend (with $\tau \approx 0.8$), indicating convergence toward a stable ordering. In contrast, *Epic-Tents* and *Meccano* display a slower upward trend, suggesting that rankings may still shift with additional annotations.

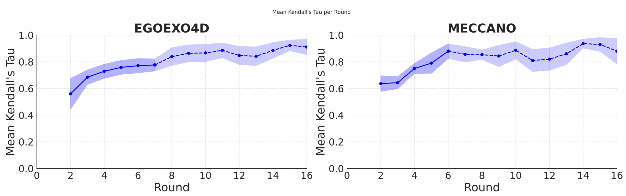
To further assess the effect of extending the annotation to more rounds, we extended beyond 6 rounds for the repre-

	Ikea [3]	Meccano [28]	Assembly101 [30]	Egoexo4D [12]	EpicTent [14]	Total
Total Clips	160	80	560	191	144	1135
Total Actions	10	5	35	12	9	71
Total Time (hr)	1.28	1.06	5.49	4.70	1.59	14.12
AVG $\pm$ STD (s)	28.88 $\pm$ 19.69	47.59 $\pm$ 21.45	35.3 $\pm$ 25.27	88.14 $\pm$ 90.93	39.71 $\pm$ 34.18	44.75 $\pm$ 48.46
Train Set (#clips)	100	48	350	119	85	702
Test Set (#clips)	40	19	140	48	36	283
Validation Set (#clips)	20	13	70	24	23	150

Table 2. Statistics of the PROSKILL dataset.



(a) Kendall's τ between consecutive rounds. A clear stabilization trend is observed for both IKEA Assembly and EgoExo4D, indicating convergence of absolute rankings over time.



(b) Comparison of ranking stability across rounds, when extending rounds beyond 6, for Meccano and EgoExo4D. Meccano exhibits earlier convergence, while EgoExo4D stabilizes more gradually.

Figure 3. Ranking stability analysis across datasets. (a) Temporal evolution of Kendall's τ across consecutive rounds. (b) Comparative round-level stability, highlighting convergence behavior.

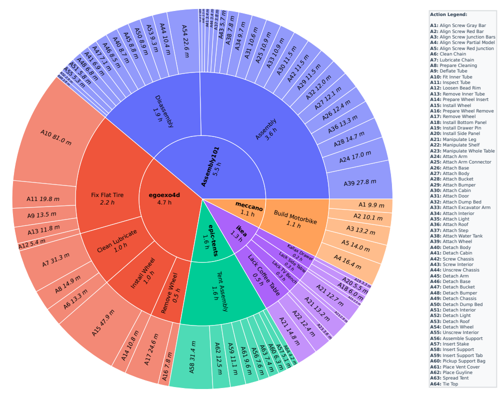


Figure 4. Illustration of the hierarchical structure of the ProSkill dataset, decomposing total annotation time by dataset, overall task type, and individual actions. Segment sizes are proportional to cumulative durations, highlighting dominant activities across different tasks. Action labels are abbreviated for readability, with a side legend offering full descriptions.

sentative datasets, *EgoExo4D* and *Meccano*, to a full round-robin scheme (i.e., all possible segment pairings). Results are shown in Figure 3b. Coherently with our previous observation, the Kendall's τ correlation between consecutive rounds stabilizes around round 6 for *Meccano*. In *EgoExo4D*, although rankings also appear stable around round 6, continuing the annotation process to round-robin yields a moderate further increase in Kendall's τ (about 6%), but also leading to a much larger number of comparisons ($\approx$ 60% more).

Data Statistics PROSKILL is composed of a total of 1135 clips and 71 actions. This makes it a large and diverse dataset for skill assessment. Action numbers, length, dataset spits are summarized in Table 2. Figure 4 summarizes the distribution of video length across datasets, tasks, and actions to illustrate the diversity and real-world quality of PROSKILL. Differently from previous datasets, we cover a wide range of activities and naturally obtain an unbalanced dataset, where the length of actions is affected by the chosen activity. Figure 5 reports ranked clips from a representative action (*Mount the table legs*) in the *IkeaASM* dataset. The figure showcases three clips positioned at different skill levels, top-ranked, mid-ranked, and low-ranked, according to their final ELO-based score. The top-ranked clip features a highly committed performer who skillfully attaches two legs simultaneously, one with each hand, demonstrating both efficiency and confidence. The mid-ranked clip shows a participant performing the task slowly, occasionally pausing to speak, indicating a less focused execution. The lowest-ranked clip reveals an external disturbance, a child interfering with a table leg during the procedure, affecting the fluency of the action. See the supplementary material for additional qualitative examples.

To ensure a robust and unbiased evaluation, the dataset was split into training, validation, and test sets at the video level. Videos were then assigned to the test, training, or validation sets such that each split contained a predefined number of clips per action, and no video appeared in more than one split. Split sizes are summarized in Table 2.

5. Experiments

PROSKILL enables for the first time a comprehensive benchmark of state-of-the-art approaches for video-based procedural skill assessment. In this section, we establish baselines to provide a clear measure of the current capabil-

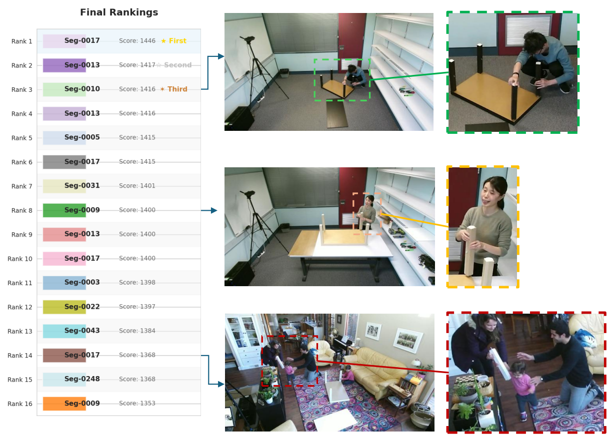


Figure 5. Example of skill ranking for the action *Mount the table legs* from the IkeaASM dataset. Clips show examples of high (green), medium (yellow), and low (red) ELO-based scores, illustrating differences in efficiency, focus, and interruptions.

ities and limitations in the field.

Problem Formulation We tackle skill assessment from video under two complementary setups: a *regression* setting, where a model predicts a continuous score $\hat{s}_i$ for each performance i, and a *pairwise ranking* setting, where the goal is to infer the relative ordering between two performances (i, j). For regression, evaluation is based on Spearman's rank correlation coefficient ρ between predicted and ground-truth scores, assessing the preservation of relative skill order. In the pairwise setting, we compute the winner prediction accuracy, i.e., the proportion of correctly predicted preferences.

Baselines For Global Ranking, we compare three representative approaches: USDL [31], DAE-AQA [40], and CoFInAl [42]. For USDL [31], we adopt the single-judge variant, as our scores, although derived from multiple annotators, are aggregated using the ELO algorithm and treated as a single ground-truth value. Both USDL [31] and DAE-AQA [40] operate on a single [1, 1024] feature vector per video, obtained by aggregating features extracted with a temporal sliding window. During training, we sample one of the T available representations for each clip. In contrast, CoFInAl [42] is designed to work with multiple temporal representations and processes four feature vectors per clip. For Pairwise Ranking we evaluate the following three representative approaches: RAAN [6], AQA-TPT [1], and CoRe [39]. The first baseline, RAAN [6], operates by comparing two input videos and predicting which one demonstrates higher skill. Importantly, both videos are sampled from the test set and were never seen during training, making this a true generalization test. RAAN [6] outputs a binary decision indicating the "winner" between the two clips, based solely on relative performance. In contrast, the other two baselines, AQA-TPT [1] and CoRe [39], perform evaluation using a reference-based approach. Each takes as input a test video and compares it against a training video depicting the same action. They estimate a score delta, reflecting how the performance in the test video deviates from that in the training example. While this method provides fine-grained feedback, it relies on access to training examples during inference, limiting its generalization compared to RAAN [6].

Implementation Details The considered baselines involve several hyperparameters that strongly influence performance. For fair comparisons, we adopt a unified feature extraction pipeline and tune hyperparameters as detailed below. Each video is split into overlapping 16-frame clips, resized to 256 px, center-cropped to 224×224, and encoded with either I3D [4] or VideoMAE [32], yielding 1024-D features. Unlike prior datasets, PROSKILL covers a wide range of actions, making per-action models impractical. We therefore train a single model across all actions for each subset (Meccano, Egoexo4D, Ikea, Assembly101, EpicTent). During training, to maximize validation performance, we optimize three hyperparameters per model: score normalization (except for RAAN), noise augmentation, and learning rate. Full grid-search details are reported in the supplementary material.

Global Ranking Results Table 3 compares Spearman's ρ across five datasets using global ranking methods (USDL, DAE-AQA, CoFInAl) and pairwise regression models (AQA-TPT, CoRe). Global ranking models generally outperform pairwise ones, suggesting that leveraging global context improves skill assessment. COFINAL achieves the highest correlations overall, notably on *Meccano* ($\rho = 0.59$), *Ikea*, and *EpicTent*, demonstrating strong temporal modeling. USDL shows consistent results, especially on *Egoexo4D*, while DAE-AQA is less stable. Among pairwise models, AQA-TPT performs moderately on *Egoexo4D* but poorly elsewhere; CORE has mixed results with some strengths on *Assembly101* and *Egoexo4D*. VideoMAE features generally lead to better correlations than I3D, except on *Assembly101* where I3D is slightly better, indicating domain-specific differences. Dataset-wise, *Meccano* is the easiest to predict, likely due to its controlled setting, while *Assembly101* is the most challenging due to high variability and noisy annotations. Overall, results confirm that skill assessment in complex, multi-action tasks remains a challenging problem requiring improved models and features.

Single-Action Model vs Multi-Action Model Table 4 compares USDL when training a unified model for all actions in a dataset or a single model per action, a setup commonly used in AQA approaches. Results show that unified

Method	Features	Ikea [3]	Meccano [28]	Assembly101 [30]	Egoexo4D [12]	EpicTent [14]
USDL [31]	I3D	0.12	0.38	0.12	0.33	0.17
	VideoMAE	0.19	**0.43**	0.13	0.39	0.23
DAE-AQA [40]	I3D	0.20	0.24	0.20	0.16	0.23
	VideoMAE	0.10	**0.42**	0.03	0.33	0.26
CoFInAL [42]	I3D	0.26	**0.59**	0.14	0.20	0.21
	VideoMAE	0.26	0.31	0.11	0.28	0.23
AQA-TPT [1]	I3D	0.14	0.12	-0.02	0.17	-0.01
	VideoMAE	0.21	0.35	0.15	**0.36**	-0.01
CoRe [39]	I3D	0.22	-0.12	0.22	0.33	0.04
	VideoMAE	0.19	0.24	0.06	**0.35**	0.12
Average	I3D	0.20	0.24	0.13	0.24	0.13
	VideoMAE	0.20	0.35	0.10	0.34	0.17

Table 3. Spearman's ρ for global ranking. A single model is trained for all the actions. Evaluated with I3D and VideoMAE features. White backgroud rows use absolute ranking, gray background uses pairwise ranking. Bold best per method, Underline best per Dataset.

Method	Features	Ikea [3]	Meccano [28]	Assembly101 [30]	Egoexo4D [12]	EpicTent [14]
USDL [31]	I3D	0.12	0.38	0.12	0.33	0.17
	VideoMAE	0.19	0.43	0.13	0.39	0.23
USDL - Single [31]	I3D	-0.18	0.09	-0.09	0.19	0.17
	VideoMAE	0.08	0.01	-0.31	0.04	0.06

Table 4. Spearman's ρ for global ranking training a model for each sub-actions.

Method	Features	Ikea [3]	Meccano [28]	Assembly101 [30]	Egoexo4D [12]	EpicTent [14]
USDL [31]	I3D	0.19	0.38	0.12	0.33	0.17
	VideoMAE	0.22	0.43	0.13	0.39	0.23
USDL [31] + Grounding	I3D+MiniLM	0.24	0.36	0.12	0.33	0.20
	VideoMAE+MiniLM	0.27	0.50	0.13	0.41	0.18

Table 5. Spearman's ρ for global ranking. A single model is trained for all the actions considering also action text embeddings. Evaluated with I3D and VideoMAE features. Underline best per Dataset.

models consistently outperform per-action models, which often suffer from poor or even negative correlations due to limited supervision at the action level. This highlights how in a realistic scenario, as the one offered by ProSkill, single-action models obtain limited performance.

Textual Grounding While training action-specific models is unfeasible in real settings, the knowledge of the action to be assessed may still be available at test time, hence constituting a useful prior for skill prediction. To assess this hypothesis, we propose to condition the unified model on a textual representation of the action (e.g., the action name) to enable it to adapt its scoring to the specific semantic context. Specifically, we conduct experiments with USDL by combining video features with textual embeddings extracted using a MiniLM pre-trained model [35]. As shown in Table 5, conditioning the unified model on action descriptions provides modest but consistent improvements in several datasets, suggesting that textual grounding can offer helpful contextual information.

Pairwise Ranking Results Table 6 shows pairwise ranking accuracy results across datasets and features. On average VideoMAE has better results over I3D across all subsets, except *Ego-Exo4D*, according to this evaluation measure. The best result of all experiments is obtained using

Method	Features	Ikea [3]	Meccano [28]	Assembly101 [30]	Egoexo4D [12]	EpicTent [14]
AQA-TPT [1]	I3D	0.71	0.66	0.53	0.72	0.70
	VideoMAE	0.73	0.74	0.70	**0.79**	0.76
CoRe [39]	I3D	0.73	0.62	0.69	**0.76**	0.67
	VideoMAE	0.70	0.73	0.69	0.75	0.75
RAAN [6]	I3D	0.45	0.52	0.51	0.58	0.52
	VideoMAE	**0.68**	0.52	0.51	0.46	0.57
Average	I3D	0.63	0.60	0.57	0.68	0.63
	VideoMAE	0.70	0.66	0.63	0.66	0.69

Table 6. Pairwise ranking accuracy. In AQA-TPT [1] and CoRe [39] one video is from test the other is an example from train, in RAAN [6] both are from test. Bold best per method, Underline best per Dataset.

AQA-TPT with VideoMAE on *Ego-Exo4D* 0.79, whereas the worst result is obtained by RAAN using I3D on *Ikea* subset 0.45. *Assembly101* confirms to be the most challenging subset also in the pairwise ranking scenario: averaging across methods and features it reach an accuracy of 0.60, just above chance level. *Meccano* is challenging as well, reaching low average scores of 0.63 and 0.6 depending on the choice of features. Overall, these results confirm that performing skill assessment in procedural video is challenging, suggesting that ProSkill is a useful resource to advance the field.

6. Conclusions

We introduced ProSkill, the first dataset specifically designed for action-level skill assessment in procedural tasks. Differently from previous datasets, ProSkill offers both pairwise and absolute ranking annotations, collected through a scalable ELO-based annotation protocol that combines the efficiency of crowdsourcing with the reliability of tournament-style comparisons. Our experiments show that existing state-of-the-art methods, originally developed for more constrained or domain-specific settings, tend to underperform on ProSkill, particularly in global ranking tasks. Pairwise methods generally yield better results, but the overall performance indicates that skill assessment in procedural videos remains a challenging and largely unsolved problem. ProSkill serves as a valuable benchmark for the evaluation of models under realistic conditions, and for investigation of the numerous open questions in human performance analysis. We will make the dataset and annotation protocol publicly available for academic research.

Acknowledgements

This research is supported by Toyota Motor Europe, Next Vision s.r.l. and by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006.

References

[1] 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*, pages 422–438. Springer, 2022. 4, 7, 8

[2] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[3] Yizhak Ben-Shabat, Xin Yu, Fatemeh Sadat Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The IKEA ASM Dataset: Understanding People Assembling Furniture through Actions, Objects and Pose. May 2023. arXiv:2007.00394 [cs]. 2, 5, 6, 8

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[5] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6057–6066, 2018. 2, 3, 4

[6] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7862–7871, 2019. 1, 2, 3, 4, 7, 8

[7] Arpad E. Elo. *The Rating of Chess Players, Past and Present*. Arco Publishing, New York, 1978. 1, 2, 5

[8] Shijia Feng, Michael Wray, Brian Sullivan, Youngkyoon Jang, Casimir Ludwig, Iain Gilchrist, and Walterio Mayol-Cuevas. Are you struggling? dataset and baselines for struggle determination in assembly videos, 2024. 2

[9] Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, and Fabio Galasso. Prego: online mistake detection in procedural egocentric videos, 2024. 1

[10] Pascal Führlich, Ágnes Cseh, and Pascal Lenzner. Improving ranking quality and fairness in swiss-system chess tournaments. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1101–1102, 2022. 5

[11] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3, 2014. 2, 3

[12] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, and others. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1, 2, 3, 5, 6, 8

[13] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 1, 2, 3

[14] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epictent: An egocentric video dataset for camping tent assembly. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3, 5, 6, 8

[15] 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*, pages 1–10, 2023. 2

[16] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 691–699, 2018. 1

[17] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment, 2021. 1

[18] J A Martin, G Regehr, R Reznick, H Macrae, J Murnaghan, C Hutchison, and M Brown. Objective structured assessment of technical skill (osats) for surgical residents. *British Journal of Surgery*, 84(2):273–278, 02 1997. 2

[19] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8310–8320, June 2025. 1

[20] 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*, pages 3204–3213, 2024. 3

[21] Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 105–123. Springer, 2022. 1

[22] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1468–1476. IEEE, 2019. 2, 3

[23] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events, 2017. 3

[24] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *2021 IEEE 23rd international work-*

shop on multimedia signal processing (MMSP), pages 1–5. IEEE, 2021. 2, 3

[25] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017. 2, 3

[26] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 1, 2, 3

[27] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 556–571, Cham, 2014. Springer International Publishing. 2, 3

[28] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021. 2, 5, 6, 8

[29] Luigi Seminara, Giovanni Maria Farinella, and Antonino Furnari. Differentiable task graph learning: Procedural activity representation and online mistake detection from egocentric videos, 2025. 1

[30] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *CVPR 2022*, 2022. 1, 2, 3, 5, 6, 8

[31] 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)*, pages 9836–9845, Los Alamitos, CA, USA, June 2020. IEEE Computer Society. 3, 7, 8

[32] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 7

[33] Aleksandar Vakanski, Hyung-Pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. *Data*, 3:2, 01 2018. 1

[34] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 4902–4910, 2021. 3

[35] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. 8

[36] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20213–20224, Paris, France, Oct. 2023. IEEE. 3

[37] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12):4578–4590, 2019. 1, 2, 3

[38] 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*, pages 2949–2958, 2022. 1, 2

[39] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7919–7928, 2021. 4, 7, 8

[40] Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. Auto-encoding score distribution regression for action quality assessment. *Neural Comput. Appl.*, 36(2):929–942, Oct. 2023. 3, 7, 8

[41] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2405–2414, 06 2023. 2

[42] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: enhancing action quality assessment with coarse-to-fine instruction alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. 3, 7, 8