

# BASKET 🏀: A Large-Scale Video Dataset for Fine-Grained Skill Estimation

Yulu Pan  
 UNC Chapel Hill  
 yulupan@cs.unc.edu

Ce Zhang  
 UNC Chapel Hill  
 cezhang@cs.unc.edu

Gedas Bertasius  
 UNC Chapel Hill  
 gedas@cs.unc.edu

<https://sites.google.com/cs.unc.edu/basket>

## Abstract

We present BASKET, a large-scale basketball video dataset for fine-grained skill estimation. BASKET contains 4,477 hours of video capturing 32,232 basketball players from all over the world. Compared to prior skill estimation datasets, our dataset includes a massive number of skilled participants with unprecedented diversity in terms of gender, age, skill level, geographical location, etc. BASKET includes 20 fine-grained basketball skills, challenging modern video recognition models to capture the intricate nuances of player skill through in-depth video analysis. Given a long highlight video (8-10 minutes) of a particular player, the model needs to predict the skill level (e.g., excellent, good, average, fair, poor) for each of the 20 basketball skills. Our empirical analysis reveals that the current state-of-the-art video models struggle with this task, significantly lagging behind the human baseline. We believe that BASKET could be a useful resource for developing new video models with advanced long-range, fine-grained recognition capabilities. In addition, we hope that our dataset will be useful for domain-specific applications such as fair basketball scouting, personalized player development, and many others. Dataset and code are available at <https://github.com/yulupan00/BASKET>.

## 1. Introduction

From art painting with creativity to scoring a three-point shot with precision, we observe and admire human skills in many domains. People in various fields strive to master skills to continuously push the boundaries of our bodies, minds, and the world. Subsequently, others are drawn to watch these people demonstrate their exceptional skills for entertainment and also as a source of inspiration to improve themselves. For example, a sport like basketball, which requires many different skills, attracts 400 million fans worldwide. AI tools enabling a deeper comprehension of such fine-grained skills could lead to many practical applications,

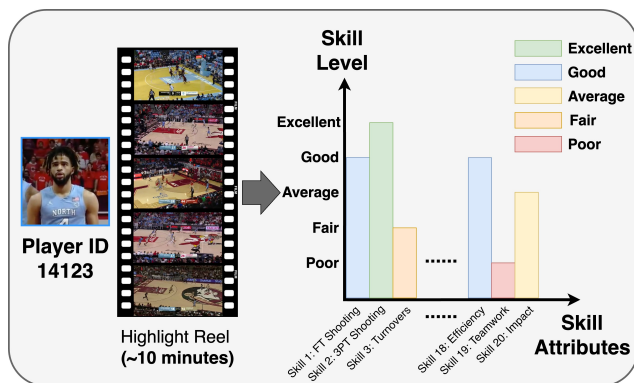


Figure 1. An illustration of our fine-grained skill estimation task. Given a long highlight video (8-10 minutes in length) that captures many plays of a particular player, the model needs to predict the skill level for 20 fine-grained basketball skills (e.g., three-point shooting, rebounding, passing, etc.). Each skill is rated on a 5-level scale, from “Poor” to “Excellent.”

from the development of personalized coaching tools to enhancing a fan’s watching experience with an expert-level commentary and skill analysis.

Recent years have witnessed remarkable progress in video models for recognizing human activities with greater precision and finer detail [4, 12, 17, 18, 24, 36–38, 49]. The video recognition field has evolved from recognizing basic actions in short clips, like those in the Kinetics [5] dataset, to identifying more complex and subtle motions in longer contexts, such as those in EgoSchema [26]. However, despite progress in general action recognition, the fine-grained skill estimation task, which requires recognizing how well the actions are performed, has received limited attention and is still highly underexplored.

One of the main limiting factors in skill estimation is the lack of large-scale video training datasets, which have effectively fueled the progress in many other video recognition domains [2, 5, 28]. As shown in Table 1, the existing skill estimation datasets [1, 8, 30, 32, 34, 39, 41, 47] are typically very small and lack significant participant diver-

Dataset	Total Video Hours	Num. Participants	Avg. Video Length (S)	Num. Skills	Num. Domains
JIGSAWS [1]	3.5	8	120	3	1
BEST [6]	26	400	188	5	5
MIT-Dive [32]	0.1	159	2.5	1	1
MIT-Skating [32]	7.3	150	175	1	1
UNLV-Dive [31]	0.4	370	3.8	1	1
MTL-AQA [30]	1.5	1412	4.1	1	1
FineDiving [41]	3.5	3000	4.2	1	1
LOGO [47]	11	200	204	1	1
Fis-V [39]	23.6	500	170	1	1
FP-Basket [3]	10.3	48	<b>780</b>	1	1
EgoExoLearn [10]	120	747	580	8	<b>8</b>
Ego-Exo4D [8]	1286	740	156	8	<b>8</b>
<b>BASKET (Ours)</b>	<b>4477</b>	<b>32232</b>	<b>500</b>	<b>20</b>	<b>1</b>

Table 1. Comparison with existing skill estimation video datasets. Our proposed BASKET dataset significantly surpasses previous datasets in scale and diversity, with 4,477 video hours and 32,232 participants. Additionally, compared to most prior datasets, our dataset provides a larger number of fine-grained skills and includes videos with a longer average duration.

sity. Even the recently collected Ego-Exo4D [8] dataset, which took 2 years, 15 institutions, and millions of dollars, only contains 800 participants, which is insufficient to train modern, data-hungry video recognition models. As a result, most existing skill estimation models are brittle and have limited generalization capabilities.

We introduce BASKET (**B**asketball **S**kill **E**stimation over **T**ime), a large-scale basketball video dataset for advancing fine-grained skill estimation. Our dataset contains over 4,400 hours of video, capturing 32,232 basketball players from 21 basketball leagues worldwide. BASKET also offers an unprecedented level of diversity in terms of participant characteristics (gender, race, age, nationality, skill level), geographic location (4 continents and over 30 countries), and the number of fine-grained skill attributes (e.g., three-point shooting, rebounding, passing, defending, etc.). We chose basketball as our primary domain for the following four reasons. First, basketball offers huge participant diversity and lots of video data that we can use to train skill estimation models. Second, basketball involves many fine-grained skills, making the skill estimation task more challenging and interesting. Third, in basketball, most players have similar visual appearances, necessitating the models to recognize fine-grained visual/skill cues rather than scene/background biases, as is common in traditional video recognition datasets. Lastly, the skilled basketball activities of each player are captured across many basketball games over several months, necessitating temporal video understanding, which many modern video modes struggle with. As shown in Table 1, our dataset significantly exceeds prior skill estimation datasets in terms of scale and variety of skills represented.

We formulate fine-grained skill estimation as a multi-way video classification problem. Specifically, given a long

8-10 minute highlight video that captures many plays of a particular player, we aim to estimate that player’s skill level over 20 fine-grained skill attributes, such as three-point shooting, passing, rebounding, defending, and many others (discussed in Section 3). Each skill is categorized into five levels: “Excellent,” “Good,” “Average,” “Fair,” and “Poor.” This is a challenging problem for several reasons. First, it requires video models to process long-form video inputs of 8-10 minutes, which is difficult for modern video recognition models. Second, due to the difficulty of obtaining bounding-box player annotations, the model has to jointly learn to identify the player of interest in each video and estimate that player’s skill. Lastly, compared to traditional object or action recognition tasks, our skill estimation task focuses on a higher-level understanding of a person’s skill, thus requiring fundamentally different visual representations that can capture subtle and nuanced skill cues rather than spatial or background cues needed for coarse action recognition [5, 9, 15, 35].

Our empirical experiments with various video recognition models, including Internvideo2 [37], VideoMamba [18], X-CLIP [24], and others, reveal that these models struggle to achieve good results on our BASKET benchmark. Specifically, we report that the best performing VideoMamba [18] only achieves **28.50%** accuracy (compared to 20% random baseline performance). In contrast, the most experienced human annotators can achieve as high as **72%** on this challenging task, highlighting the gap in fine-grained video recognition models’ capabilities. We hope that our work will encourage the skill estimation community to use our large-scale video benchmark to develop new and more powerful fine-grained skill estimation models.

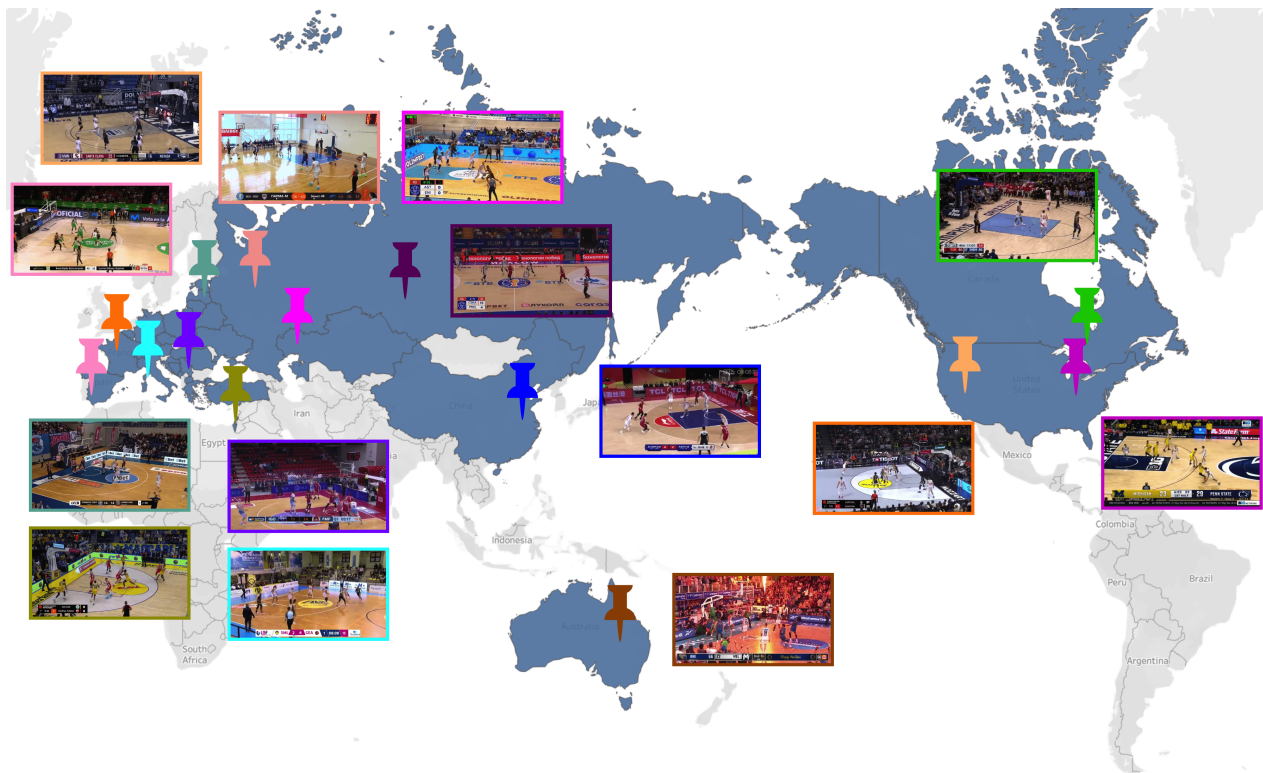


Figure 2. BASKET is a large-scale video dataset containing 4,477 hours of video and capturing 32,232 basketball players from 21 basketball leagues worldwide. Here, we showcase geographic location diversity of our dataset, i.e., it captures basketball players from 4 continents and more than 30 countries. Each pin marks the approximate geographic location of the visualized basketball game (the color of a pin corresponding to the border of the visualized game).

## 2. Related Work

**Skill Estimation Datasets.** Skill estimation has recently gained interest, particularly in domains where understanding and assessing human performance is essential. Such tasks and datasets are more challenging than general video recognition datasets such as Kinetics [5], AVA [9], UCF-101 [35], and HMDB51 [15]. In sports, most existing skill-related datasets focus on diving actions. MIT-Dive [32], UNLV-Dive [31], MTL-AQA [30], and Fine-Diving [41] are diving datasets with fine-grained annotations of action procedures, accompanied by the official judge score. LOGO [47] dataset consists of multi-person long-form artistic swimming competition videos annotated with human assessment scores. Other sports-based skill datasets are related to basketball [3], figure skating [32], and golf [27]. Beyond sports, several datasets have been collected to evaluate surgical and daily skills. JIGSAWS [1] is a video dataset designed to evaluate the surgical skills of three procedures. BEST [6] contains videos across five daily tasks and uses pairwise rankings for skill assessment tasks. EgoExoLearn [10] includes egocentric and demonstration videos to estimate skill proficiency for daily and lab tasks. Most recently, Ego-Exo4D [8] is a multi-view ego-

centric and exocentric proficiency estimation benchmark covering 8 physical and procedural scenarios. Unlike these prior datasets, our BASKET dataset offers a much greater scale, number of skills, and participant diversity.

**Skill Estimation Methods.** Several existing methods tackle action quality assessment and skill estimation tasks. The work in [14] proposed a CNN-based model to segment kinematic data into hierarchical features for surgical skill classification. Subsequently, a multi-path framework [21] was developed to combine various skill aspects from surgical videos. Building on this, the method in [44] proposed a contrastive regression framework integrated with group-aware regression to assess surgical proficiency. For sports, MTL-AQA [30] incorporates a multitask learning framework with spatiotemporal features extracted by 3D CNNs for diving quality assessment. FineDiving [41] proposes a procedure-aware action quality assessment approach that uses a temporal segmentation attention module to analyze spatial, semantic, and temporal correspondences in diving. LOGO [47] introduces a group-aware attention module that integrates spatial-temporal group dynamics to model relations between artistic swimmers in a scene. RICA<sup>2</sup> [25] uses a deep probabilistic model that integrates human score rubrics and models prediction uncertainty for action qual-



Figure 3. Our BASKET dataset covers five coarse basketball skill categories and twenty fine-grained skills, focusing on the evaluation of multi-faceted skill understanding of basketball players.

ity assessment in diving and surgical skills. NS-AQA [29] introduces a hierarchical neuro-symbolic approach for evaluating diving quality. Compared to these prior works, our main objective is not to propose a novel skill estimation method but to introduce a new, challenging, large-scale, fine-grained skill estimation dataset.

**Video Recognition Models.** Traditional video recognition models [4, 17, 18, 36] are primarily designed to recognize coarse action classes in short video clips such as Kinetics [5]. More recent work explores fine-grained video action recognition, such as recognizing gymnastic actions [19, 34], diving procedures [30, 41], basketball moves [42, 43], and figure skating [22, 23]. Furthermore, models such as Temporal Query Network [46] and PoseConv3D [7] have been proposed to address the challenges in detailed temporal video understanding. In addition to fine-grained video recognition, recent years have witnessed many methods for long video modeling, including MeMVit [38], TimeCeption [11], ViS4mer [13], VideoGraph [12], AdaptFocus [49], and VideoMamba [18]. Finally, recent video models [20, 24, 33, 37, 40, 48] have focused on learning powerful video representations from multimodal video-text data. Compared to these prior video approaches, in this work, we introduce a new, challenging fine-grained skill estimation dataset and show that modern video recognition models lack the capabilities needed to excel at it.

### 3. The BASKET Dataset

Here, we discuss the construction and characteristics of our new large-scale skill estimation video dataset, BASKET.

#### 3.1. Dataset Construction

**Collecting Full-Game Videos.** We use a basketball game replay archive to collect basketball videos across many leagues worldwide. This results in roughly 66,000 basketball games. Each full-game video also contains a transcript annotated by experts with timestamped player-event instances (e.g., At 3:03 in the first quarter, Steph Curry makes a three-point shot).

**Player Highlight Video Generation.** To generate player highlight videos, we first curate a list of 32,232 players from 21 unique basketball leagues worldwide. Since the players’ skills can vary from year to year, we treat the same players in different seasons as different subjects. We use the aligned full-game video footage and timestamped player-event instances (described above) to extract basketball events associated with each player. Since the number of events per player can be large, we randomly select 50 events for each player and then extract video clips around the selected events’ timestamps. Each event clip is, on average, 10 seconds long to capture the preceding context and the outcome of the player’s actions. Finally, the selected clips are shuffled and combined into a single highlight video of approximately 9 minutes in length. All videos are 704x400 in resolution and 30 FPS.

**Player Skill Annotations.** We also obtain expert-annotated ranks of the players for 20 fine-grained basketball skills, such as three-point shooting, passing, rebounding, and others (see Section 3.2 for a more detailed analysis). The skill level labels are obtained by the basketball experts who review the game footage and annotate all the events in the game. These annotations are then aggregated and used to rank each player in 20 skill categories. To divide all players into five skill level categories (i.e., “Excellent,” “Good,” “Average,” “Fair,” and “Poor”), we sort the ranks of all players within the same league same season from highest to lowest and divide them evenly into 5 skill levels. We believe that five skill level categories make the task sufficiently challenging but still solvable. We also observe that different leagues (e.g., college vs. professional) may exhibit significant skill level differences. To account for these differences and avoid player comparisons across different leagues, we construct the skill labels separately for each league and season. During model training/inference, the model has to implicitly learn to predict skill levels specific to each basketball league.

#### 3.2. Dataset Statistics

BASKET consists of 32,232 player videos with an average duration of 500 seconds. 7,563 of these videos include women players. The dataset encompasses 21 basketball leagues from over 30 countries across 4 continents (See Figure 2) over 6 seasons, from 2017 to 2023. The dataset also represents players with a broad range of experience and age,

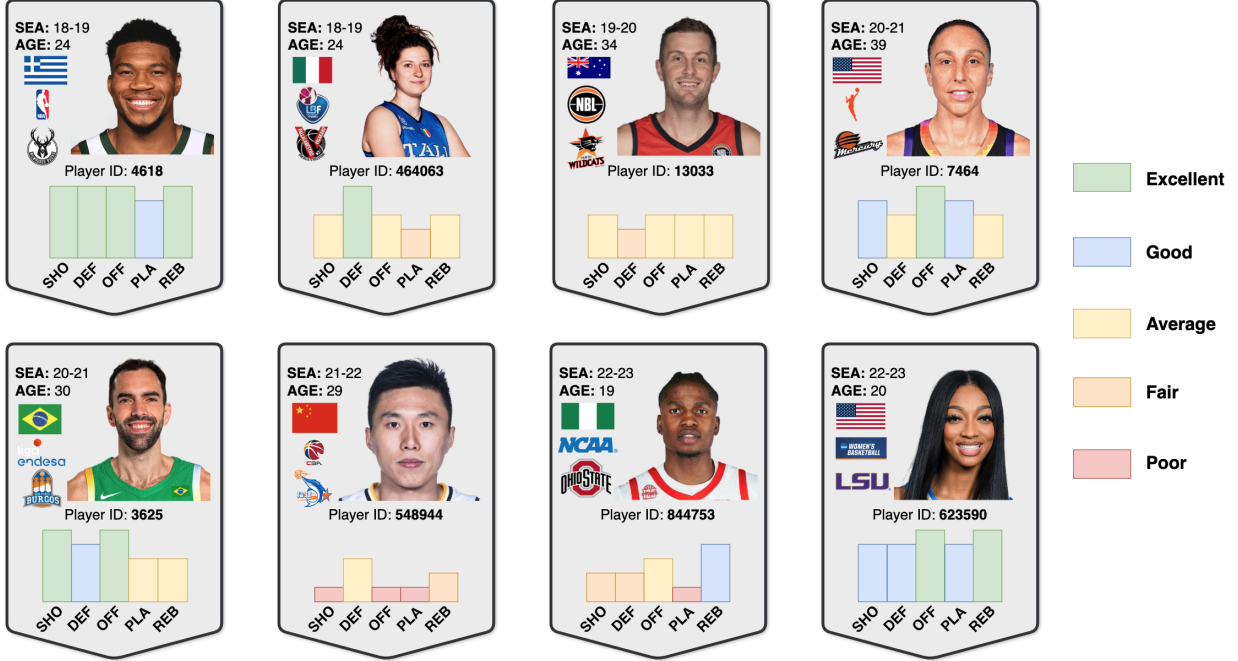


Figure 4. Visualizing some of the players from BASKET dataset. Our dataset offers unprecedented player diversity in terms of player nationality, age, gender, race, experience, and skill. The left side of each profile card displays the season, player nationality, league, and club. Skill levels are derived by averaging finer-level skills within each coarse category (as described in Section 3). **SHO**: Shooting, **DEF**: Defense, **OFF**: Offense, **PLA**: Playmaking, **REB**: Rebounding

from college-level athletes to professional players with 10-20 years of experience. Figure 4 showcases some of the players from our BASKET dataset, highlighting diversity in nationality, age, gender, race, experience, and skill.

As shown in Figure 3, BASKET encompasses 20 fine-grained basketball skills. These skills can be divided into five broader categories, including shooting, playmaking, defense, rebounding, and offense. These coarse categories can be further divided into finer-grained skills (See Figure 3).

#### 4. Fine-Grained Skill Estimation Task

Given a long player highlight video, we aim to classify each of the 20 basketball skills into 5 levels from “Poor” to “Excellent.” Formally, given a video  $V = \{x_t\}_{t=1}^T \in \mathbb{R}^{3 \times T \times H \times W}$  with  $T$  frames, where each frame  $x_t \in \mathbb{R}^{3 \times H \times W}$  represents an RGB image of size  $H \times W$ , the model needs to output a set of skill level predictions  $Y = \{y_s\}_{s=1}^S$  where  $S$  is the number of skills (i.e., 20 in our setting). Each prediction  $y_s \in \mathbb{R}^L$  represents the probabilities for each skill level, and  $L$  is the number of skill levels (i.e., 5 in our setting).

This task presents several key challenges. First, the video model must be able to handle long video inputs. This includes long-term reasoning capabilities about different short-term segments within the video, as well as efficient long-video processing to enable scalable training on

many long videos. Second, since the video inputs do not contain bounding box annotations for the player of interest, the model needs to jointly identify (implicitly) the recurring player (i.e., the player of interest) in all the video segments and then estimate that player’s skills. This can be very challenging, as basketball videos often contain scenes with fast-moving motions, heavy occlusions of players, and significant camera cuts. Third, since each player’s highlight is aggregated from only 50 event clips, the model needs to extrapolate that player’s skill level by analyzing subtle cues of player performance, such as differences in techniques, player posture/pose, efficiency/effectiveness of their actions, decision-making, team dynamics, and many other factors. Achieving this requires fundamentally different perception capabilities focusing on nuanced skill cues rather than coarse background cues as is typical with standard action recognition tasks [5, 9, 15, 35].

#### 5. Experimental Setup

In the following section, we provide details on our dataset splits, evaluation metrics, and baselines we consider.

**Dataset Splits.** We split 32,232 players from BASKET into training, validation, and test splits using the 7:1:2 ratio. Since our considered videos span seasons from 2017 to 2023, we ensure that each unique player is consistently included within the same set (i.e., all Stephen Curry’s videos

Method	Pretraining Datasets	# Frames	# Params (M)	Test Acc. (%)
Random Baseline	-	-	-	20.00
SigLIP [45]	WLI+K400	64	203	21.85
MeMVIT [38]	K400	64	212	23.01
LLaVA-OneVision [16]	CC+BLIP+SD+UR+SG	32	8200	23.84
X-CLIP [24]	WIT+K400	16	196	24.37
VideoMAE2 [36]	UH+K400	32	1012	24.43
TimeSformer [4]	IN21K+K400	96	121	25.21
UnmaskedTeacher [17]	K710+CC+VG+SBUC+CC15M+WV12M	32	90	26.97
InternVideo2 [37]	LA+WV10M+WV2M+SC+K710	32	1024	27.52
VideoMamba [18]	IN1K+K400	64	74	<b>28.50</b>

Table 2. Comparison of various video recognition models on our fine-grained skill benchmark, BASKET. All experiments were conducted with a uniform video frame sampling strategy with a 224x224 spatial video resolution by fine-tuning each model with its best configuration. These results show that none of the methods achieves over 30% accuracy, indicating a large room for future improvement. **IN**: ImageNet, **K**: Kinetics, **WLI**: WebLI, **WIT**: WebImageText, **UH**: UnlabeledHybrid, **LA**: LAION, **WV**: WebVid, **SC**: Self-collected, **CC**: COCO, **VG**: Visual Genome, **SBUC**: SBU Captions, **UR**: UReader, **SD**: SynDOG, **SG**: ShareGPT4V

are in the training set). Additionally, we exclude videos from the 2017-2018 season from the training data and use them only for testing to validate the model generalization to previously unseen season data. Furthermore, we also exclude the videos from 4 of our selected leagues (from 21 total leagues) from the training data and use it only for testing to validate the generalization to the data from the previously unseen basketball leagues. In total, we use roughly 19,500 players for training, 2,800 for validation, 5,600 for testing, and 4,500 for generalization experiments.

**Evaluation Metrics.** We use top-1 accuracy to evaluate the skill estimation performance in each skill category. To obtain a single skill estimation metric, we average the accuracies across all 20 fine-grained skill categories.

**Baselines.** Since our benchmark and task do not have well-established baselines, we implement a number of our own baselines to measure the performance on our BASKET benchmark. *TimeSformer* [4] is a convolution-free video classification model that relies solely on self-attention over space and time, adapting the Vision Transformer (ViT) for video understanding. *MeMVIT* [38] introduces a memory-augmented multiscale vision transformer for long-term video recognition, processing videos incrementally and caching the extracted information into memory for referencing past context. *VideoMAE2* [36] uses a masked video autoencoding strategy to learn powerful spatiotemporal features in a self-supervised manner. *Unmasked Teacher (UMT)* [17] learns strong video representations by aligning unmasked video tokens with the representations of foundational image models. *InternVideo2* [37] learns spatiotemporally consistent video representations via distillation from models like InternVL and VideoMAEv2, which capture spatial, temporal, and multi-modal information from the video. *VideoMamba* [18]

adopts a bidirectional selective state-space model (SSM) architecture for scalable and memory-efficient long video processing. *X-CLIP* [24] extends CLIP to video-text retrieval with multi-grained contrastive learning, aligning coarse-grained and fine-grained visual features. *SigLIP* [45] uses a pairwise sigmoid loss to learn visual representations from large-scale image-language data. To extend SigLIP to video, we apply temporal pooling on individually extracted frame-level features. *LLaVA-OneVision* [16] introduces a large vision-language model that achieves impressive cross-scenario generalization transfer across many image and video tasks. To adapt this model to our setting, we converted our dataset into text format and asked the model to output skill predictions for each category in the text format. **Implementation Details.** For all our experiments we use eight NVIDIA RTX A6000 GPUs, each with 48G of memory. All models are fine-tuned to optimal performance using the best available pre-trained checkpoints and hyperparameters. We ablate on several key hyperparameters in Section 6. Each video model uses 20 classification heads (one linear layer each) for predicting each of the 20 skill categories. For all experiments, we sample the video frames uniformly and resize the shorter side of the frame to 224 pixels.

## 6. Experimental Results

In this section, we first present the results of all the baseline models on our BASKET benchmark. Afterward, we present the human study results and the ablation studies conducted on the best-performing video model.

### 6.1. Main Results

In Table 2, we present the performance of the latest state-of-the-art video recognition models on our fine-grained skill

Season	18-19	19-20	20-21	21-22	22-23	Gender	Male	Female	Location	N. America	Europe	Asia	Australia
Acc. (%)	27.12	27.41	27.36	29.38	29.92	Acc. (%)	27.51	31.33	Acc. (%)	29.68	25.58	24.24	23.00
(a) Test Accuracy Across Seasons						(b) Test Accuracy by Gender			(c) Test Accuracy by Location				

Table 3. We study how VideoMamba generalizes to videos with different 1) seasons, 2) player gender, and 3) geographic locations.

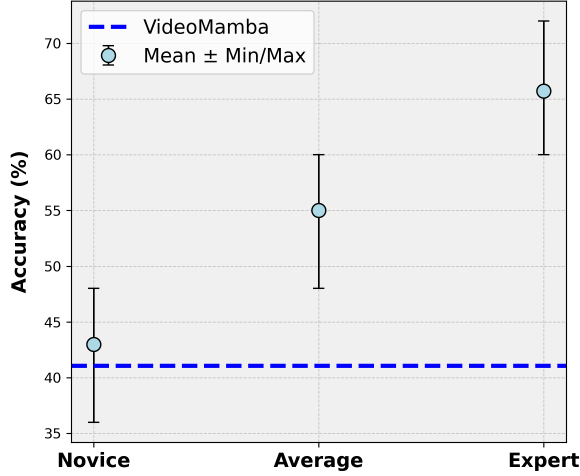


Figure 5. We visualize our human study results on BASKET, with subjects grouped by their expertise level (i.e., novice, average, expert). For each group, we visualize the mean accuracy and the min/max ranges. The blue dashed line indicates the performance of our best model, VideoMamba [18]. To ensure that the time needed to complete the study is reasonable, every subject is asked to watch videos of 5 uniformly selected players and classify 5 selected skills into 3 skill levels (i.e., “Poor,” “Average,” and “Excellent”). Our VideoMamba baseline, which was not trained on these players, is also tested in this exact setting. Our results highlight the gap between model and human performance, especially for the human subjects with high expertise.

estimation benchmark, BASKET. These results suggest that all models perform poorly, with none achieving more than 30% accuracy (the random baseline being 20%). From the results, we observe that the best-performing model is VideoMamba, likely due to its long-range temporal reasoning capability, which is crucial for capturing the extended context in our skill estimation task. Furthermore, we note that while models pretrained on large-scale image and video datasets, such as UnmaskedTeacher and Intern-Video2, show moderate improvement over other less performant models, they remain low overall, underscoring that the task cannot be solved by simply scaling the image/video data from the Web. Lastly, our results show that the state-of-the-art vision-language models, such as LLaVA-OneVision and SigLIP, are among the worst performers. These low accuracies support our earlier claims that to do well on our challenging BASKET benchmark, we need new models with fundamentally different visual recognition capabilities.

## 6.2. Human Evaluation

Next, we conduct a study to assess human performance on our challenging skill estimation BASKET benchmark. To ensure that the time needed to complete the study is reasonable, instead of using all 20 skill categories (see Figure 3), we select one fine-grained skill from each broader skill category (i.e., shooting, rebounding, defense, playmaking, and offense) to assess 5 skills in total. Additionally, our initial human study experiments revealed that classifying each skill into 5 skill levels (i.e., “Poor,” “Fair,” “Average,” “Good,” and “Excellent”) takes too long for human subjects to complete the study (i.e., more than 1 hour for a single session). Therefore, our finalized human study involves asking each human subject to evaluate 5 players across 5 skills using 3 levels for each skill.

To conduct our human study, we used videos from the 2018–2019 NCAA Division I season. During the study, we did not provide any information about the league or the players to the subjects to avoid bias. Furthermore, subjects were asked if they recognized any players before the study and were disqualified if they answered “yes”. To reduce the variance in human scores, we asked each subject to perform the study on 3 sets of different players. Additionally, for a comparison with a computer vision model, we included our best-performing VideoMamba [18] baseline, which we fine-tuned for three-level skill classification to match the setting of the human study. Note that all players in the human study were sampled from the test set to make the comparison fair.

In Figure 5, we present our human subjects’ results, which include data from 11 subjects. We categorize the results based on the subjects’ expertise level (i.e., novice, average, expert). For each group, we calculate the average performance and plot the maximum and minimum results. Based on the results, we first report that VideoMamba achieves 41% accuracy (compared to 33.3% random baseline). We also observe that human subjects in the novice group achieved 43%, barely outperforming VideoMamba. However, when considering human subjects with higher expertise, we observe that their accuracy is much higher, with the expert group averaging 66% with a top performance reaching 72%. These results suggest that human experts can solve this skill estimation task while computer vision models struggle.

## 6.3. Generalization Analysis

In Table 3, we analyze how our best-performing VideoMamba model generalizes to 1) videos across different sea-

Num. Frames	Test Acc. (%)	Sampling Rate	Test Views	Test Acc. (%)	Data Fraction (%)	Test Acc. (%)
16	23.93	32	8	22.64	25	23.67
32	25.20	64	4	25.89	50	24.89
<b>64</b>	<b>28.50</b>	128	2	26.30	75	27.76
128	27.18	<b>Uniform</b>	<b>1</b>	<b>28.50</b>	<b>Full</b>	<b>28.50</b>

(a) **Number of Frames:** We vary the number of input frames and observe that 64 frames lead to the best accuracy.

(b) **Sampling Rate:** We investigate different frame sampling rates (i.e., the interval between consecutive frames) and report that uniform sampling works the best. We use 64 frame inputs for these experiments.

(c) **Data Fraction:** We vary the amount of training data and observe that using full data leads to the best accuracy.

Table 4. We ablate different design choices of the best performing VideoMamba model. The skill estimation performance is evaluated using the top-1 accuracy metric averaged across 20 skills.

Season	League	Test Acc. (%)
Unseen	Unseen	23.53
Unseen	Seen	26.83
Seen	Unseen	23.09
Seen	Seen	28.50

Table 5. We conduct cross-season & cross-league generalization evaluations of the best-performing VideoMamba model to test its performance on various previously unseen scenarios. We observe that the model performs substantially worse on the videos from previously unseen seasons and locations, signifying issues with out-of-domain generalization.

sons, 2) videos with players of different gender, and 3) videos with different geographic locations. Our results suggest several interesting trends. First, we observe that the model is doing slightly better in videos from more recent seasons. This can be attributed to the fact that the more recent seasons have more videos. Next, we observe that the model is doing better in the videos of female players, which is somewhat surprising since our training videos contain  $3\times$  more videos of male players. This result signifies that the model generalizes reasonably well between genders. Lastly, our results indicate that the model achieves the best accuracy in videos originating from North America and the worst accuracy in videos from Australia. We hypothesize that this is because we have  $55\times$  more videos from North America than from Australia.

#### 6.4. Cross-Season & Cross-League Generalization

Next, in Table 5, we perform evaluations of VideoMamba for cross-season and cross-league generalization in previously unseen settings. Specifically, we define three evaluation scenarios: 1) testing on videos from previously unseen seasons, 2) testing on videos from previously unseen geographic locations (i.e., leagues), and 3) the combination of 1) and 2). Our results in Table 5 suggest that compared to the testing accuracy obtained on the in-domain data (i.e., previously seen season and league), the results on previously unseen leagues drop substantially ( $> 4\%$ ). We also observe that generalizing across leagues is more difficult than across seasons.

#### 6.5. Ablation Studies

In Table 4, we use VideoMamba [18] to perform ablation studies on various model design choices, including 1) the number of input frames, 2) different frame sampling strategies, 3) the effect of the training data size, and 4) varying the clips included in the player highlight video.

**Number of Input Frames.** In Table 4a, we investigate the model’s performance with a different number of input frames. For these experiments, we use the uniform frame sampling strategy. Although the model can process 128 frames, we find that using 64 frames yields the best results.

**Frame Sampling Rate.** In Table 4b, we test different frame sampling rates with a fixed input of 64 frames. We also vary the number of temporal test views to cover the whole video input during inference. We observe that uniform sampling leads to the highest accuracy.

**Training on the Subset of Data.** In Table 4c, we study the skill estimation performance as a function of the training data size. Our results indicate that using the full data leads to the best performance.

**Varying the Selected Player Clips.** Lastly, we study how varying the clips in the player’s highlight video affects the performance. To do this, we generated 5 different highlight videos for each player by randomly sampling 50 clips each time. We report the average accuracy of 28.44%, with a variance of 1.30, which is comparable to our reported accuracy of 28.50%. Thus, based on these results, we can conclude that the selection of clips does not dramatically impact the overall results.

#### 7. Conclusion

We introduce BASKET, a diverse, large-scale basketball video dataset for fine-grained skill estimation. Our experiments with many recent state-of-the-art video recognition models revealed that all of these models struggle to achieve high skill estimation accuracy, thus highlighting the need for video models with fundamentally new recognition capabilities. We hope that our new dataset will inspire new ideas in fine-grained skill estimation and will lead to many practical applications, from promoting fair basketball scouting to developing personalized player development tools.

## References

- [1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021.
- [3] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2177–2185, 2017.
- [4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021.
- [5] 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.
- [6] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. 2019.
- [7] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022.
- [8] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.
- [9] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.
- [10] 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.
- [11] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019.
- [12] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*, 2019.
- [13] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.
- [14] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11*, pages 214–221. Springer, 2018.
- [15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.
- [16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [17] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023.
- [18] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.
- [19] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545, 2021.
- [20] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.
- [21] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9522–9531, 2021.
- [22] Shenlan Liu, Xiang Liu, Gao Huang, Lin Feng, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Hong Qiao. Fsd-10: a dataset for competitive sports content analysis. *arXiv preprint arXiv:2002.03312*, 2020.
- [23] Shenglan Liu, Aibin Zhang, Yunheng Li, Jian Zhou, Li Xu, Zhuben Dong, and Renhao Zhang. Temporal segmentation of fine-grained semantic action: A motion-centered figure skating dataset. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2163–2171, 2021.
- [24] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings*

- of the 30th ACM International Conference on Multimedia, pages 638–647, 2022.
- [25] Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. Rica<sup>2</sup>: Rubric-informed, calibrated assessment of actions. *arXiv preprint arXiv:2408.02138*, 2024.
  - [26] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
  - [27] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
  - [28] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.
  - [29] 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.
  - [30] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019.
  - [31] 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.
  - [32] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014.
  - [33] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.
  - [34] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020.
  - [35] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
  - [36] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.
  - [37] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.
  - [38] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.
  - [39] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yungang 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.
  - [40] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
  - [41] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2949–2958, 2022.
  - [42] Jinglin Xu, Guohao Zhao, Sibao Yin, Wenhao Zhou, and Yuxin Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21773–21782, 2024.
  - [43] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 208–224. Springer, 2020.
  - [44] 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 (ICCV)*, pages 7919–7928, 2021.
  - [45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
  - [46] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021.
  - [47] 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, 2023.
  - [48] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large lan-

guage models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.

- [49] Jiaming Zhou, Hanjun Li, Kun-Yu Lin, and Junwei Liang. Adafocus: Towards end-to-end weakly supervised learning for long-video action understanding. *arXiv preprint arXiv:2311.17118*, 2023.