Supplementary Material: Sports Video Analysis on Large-Scale Data

Dekun Wu*[1], He Zhao*[2], Xingce Bao[3], and Richard P. Wildes[2]
[1]University of Pittsburgh, [2]York University
[3]École Polytechnique Fédérale de Lausanne (EPFL)
dew104@pitt.edu, {zhufl, wildes}@cse.yorku.ca,
xingce.bao@alumni.epfl.ch

1 Summary

This document provides additional material that is supplemental to our main submission. Two additional human evaluation experiments for video captioning as well as for dataset quality are provided in Sections 2 & 3, respectively. In Section 4, we provide a concrete example showing where we collect the data. Sections 5 and 6 elaborate on details of finer feature collection and report the used codebase resources. We then provide a detailed description of baseline methods in Section 7, extended analysis on results of video captioning in Section 8 and introduce metrics used for action recognition and player identification in Section 9. Finally, we present a long table for the whole action name list of our NSVA in Section 10.

2 Human evaluation on generated texts

To additionally assess the quality of the sentences generated by our video captioning approach, we conducted a human evaluation by randomly selecting 50 videos from the test set. Based on these 50 videos, we collect caption outputs generated by four models, i.e., our full model (T+PB+BAL+BAS+PA), S3D+PB+BAL+BAS+PA, TimeSformer only and S3D only. We created a website for this assessment; see Figure 1. We ask two human annotators to assess model outputs on three aspects: Relevance (Player), Relevance (Action) and Naturalness. The definitions of these aspects are shown below:

- **Naturalness** denotes to what degree a text is natural. 1 means the lowest naturalness and 5 means the highest naturalness. To achieve the highest naturalness score, a generated text must be grammatically correct and free of repeated words.
- **Relevance (Action)** denotes to what degree a text is relevant to the ground-truth text on the level of action. 1 means the lowest relevance and 5

* Equal contribution
Corresponding author: dew104@pitt.edu

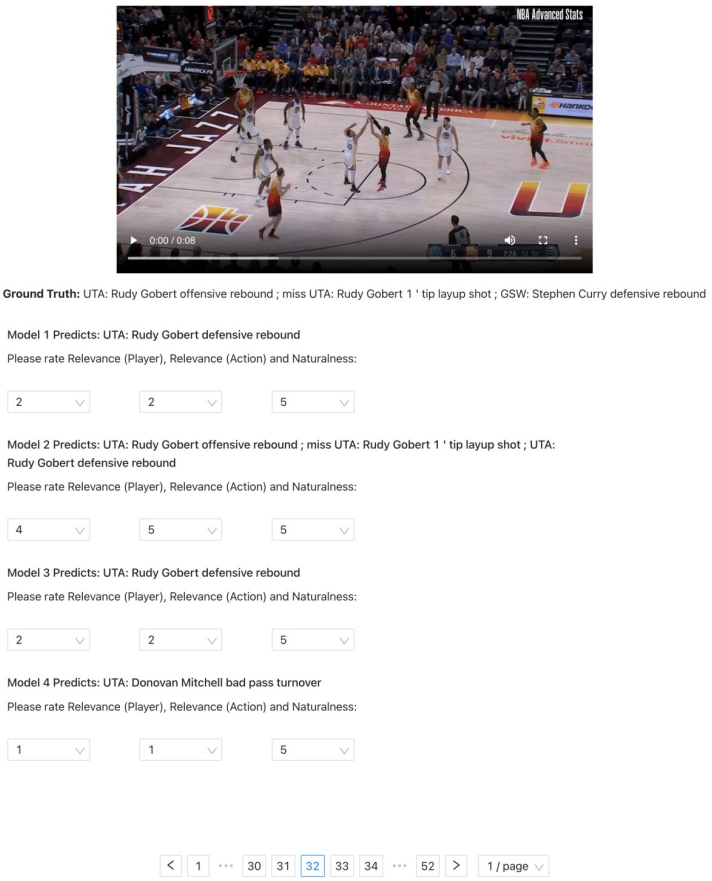

Fig. 1: A website for human evaluation on quality of generated texts (video captioning task).

means the highest relevance. To achieve the highest relevance (action) score, a generated text must describe actions correctly in the same order as the ground-truth text.

– **Relevance (Player)** denotes to what degree a text is relevant to the ground-truth text on the level of player identity. 1 means the lowest relevance and 5 means the highest relevance. To achieve the highest relevance (player) score, a generated text must describe names of players correctly in the same order as the ground-truth text.

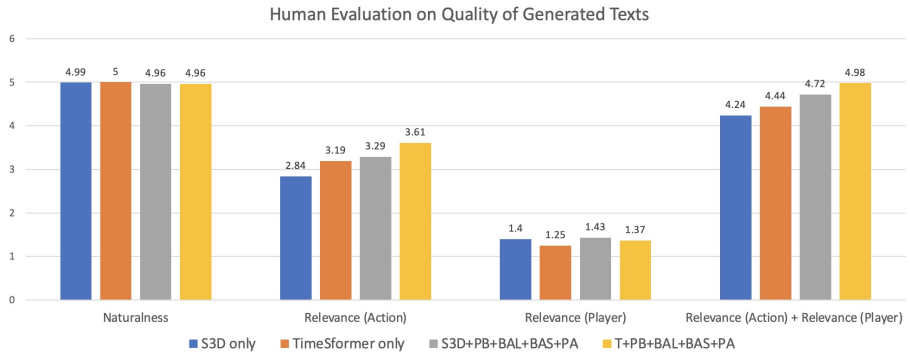

Fig. 2: Human evaluation on quality of generated texts (video captioning task).

As shown in Figure 1, we anonymize and randomize model outputs so annotators can only rely on the quality of generated texts for the assessment. We collect results from two annotators and calculate the average rating for every model output on three aspects; see results in Figure 2.

From the first group in Figure 2, we see that all models can generate natural texts. We conjecture that is because texts in our dataset are very compact, i.e., average 6.5 words per sentence. Models can easily capture the underlying pattern of texts and generate outputs following certain formats.

Results on the second group of Figure 2 show that the full model (T+PB+BAL+BAS+PA) outperforms other models by a large margin. This result is consistent with Table 4 (main paper), where we test models' outputs with automatic evaluation metrics. Results on the third group of Figure 2 show that player identification is the most challenging part of our dataset and the average score of Relevance (Player) for four models is 1.36 compared to 3.23 average score of Relevance (Action) for four models. All models struggle to generate texts with accurate player identities. The model that has the best performance on Relevance (Player) is S3D+PB+BAL+BAS+PA , but note that during the training phase all models are optimized to output the entire sentence instead of only the player identities. By combining the results of Relevance (Action) and Relevance (Player), we have results on the sentence level, which are shown in the last group of Figure 2. We see that our full model (T+PB+BAL+BAS+PA) outperforms other alternatives in human evaluation on the sentence level, which is consistent with results on automatic evaluation metrics in Table 3 (main paper).

3 Human evaluation on quality of dataset

To assess the quality of the dataset we have created, we conduct an additional human evaluation to assess 50 video-text pairs provided on the NBA official website on two aspects: Relevance and Informativeness.

- **Relevance** denotes to what degree a ground-truth text is relevant to the video. 1 means the lowest relevance and 5 means the highest relevance. To achieve the highest relevance score, the events described by a ground-truth text should be seen in its associated video.
- **Informativeness** denotes to what degree a ground-truth text covers events that happen in its associated video. 1 means the lowest informativeness and 5 means the highest informativeness. To achieve the highest informativeness score, a ground-truth text should cover all events seen in its associated video.

We create a website to enable this assessment (see Figure 3) and ask two annotators to evaluate 50 video-text pairs extracted from the test set on two aspects, i.e., relevance and informativeness. Results are shown in Figure 4.

From Figure 4, we can see that our dataset is of very high quality. 5 out of 5 Relevance score means that in every video-text pair, events that a text describes can be seen in the video. 4.6 out of 5 informativeness score means that a ground-truth text can cover most of the events happening in a video despite having some events missing. In most such cases, missing events can be found in the text of preceding or subsequent videos that partially overlap with the current one. For example, four events "Rubio miss 20' jump shot; Gobert offensive rebound; Gobert miss 1' tip layup shot; Stephen Curry defensive rebound" are present in two consecutive video-text pairs of which the second video overlapped with the first one and its text only describes the last three events seen in the video; see Figure 5. We will target this problem and solve it by combining overlapping videos with their texts for the next version of our dataset.

4 Data collection details

Our data is collected from official NBA websites. An example can be viewed from the link: https://www.nba.com/game/mil-vs-lal-0021900939/play-by-play. It is seen that an extensive amount of captions are recorded for the entire game of Milwaukee Bucks vs. Los Angeles Lakers on March 6^{th}, 2020. These captions are organized along the timelines of the play development, i.e., from the first period to the last. More importantly, these captions describe essential information, e.g., player identifications, actions and event results, which is extremely useful for statistics tracking and post-game summary. In fact, some efforts have been made to convert professional captions to other valuable assets, e.g., box scores [1], game charts [2] and whole-game summaries [3]. We believe our study can naturally support many similar downstream tasks.

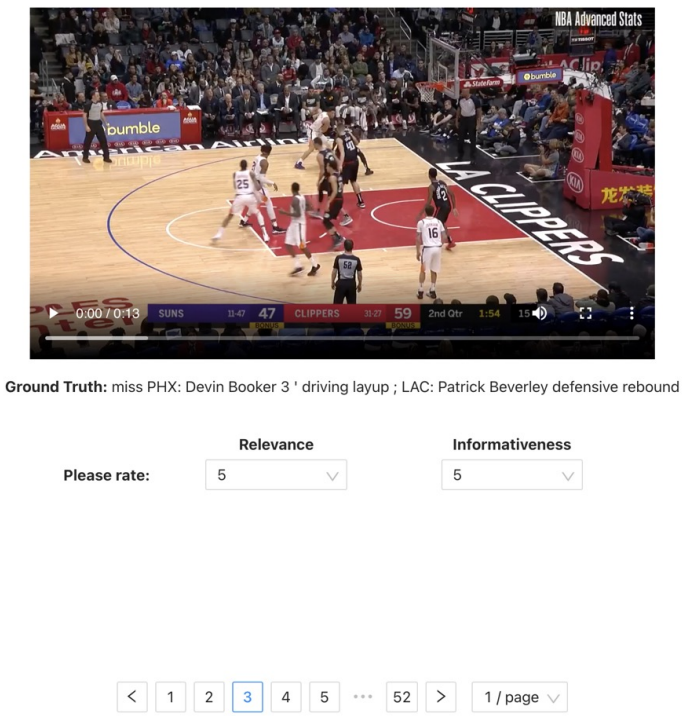

Fig. 3: A website for human evaluation on quality of our dataset.

5 Courtline segmentation

For the courtline segmentation used as input to one of our finer feature modules, we resort to existing work [13] where a pre-trained courtline segmentation network (built uon the pix2pix [7] framework) is provided. We use the code and model weights from link: https://github.com/luyangzhu/NBA-Players. A high resolution example (the same as Figure 2 in main paper) can be viewed in Figure 6 in this document. Though the segmented result (see right sub-figure) is not perfect, i.e., some areas are occluded by the score box or players, it nevertheless delineates the boundary, the penalty circle and the three-point curve.

6 Player detection

For player detection, we use the YOLOv5 code as well as MS-COCO pre-trained weights from the link: https://github.com/ultralytics/yolov5/. As for additional annotations for the ball and basket, we use the bounding box annotation

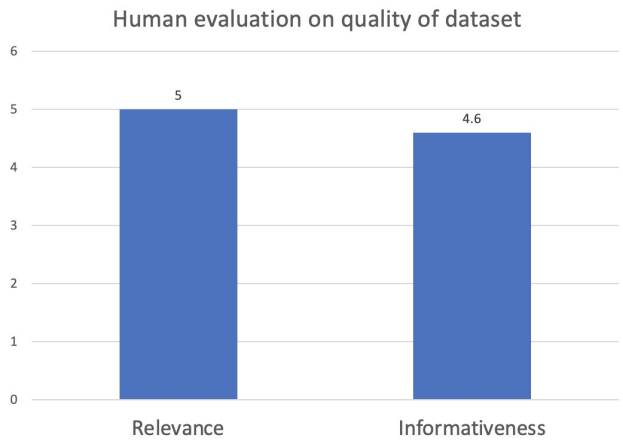

Fig. 4: Human evaluation on quality of our dataset.

tools from link: https://github.com/tzutalin/labelImg. We show a high resolution result in Figure 7. It is seen that our detection contains noisy signals, such as referees and audience. However, our approach is robust to such disturbance to some extent as we focus on the interaction between ball and players (defined as $\mathbf{I}_{pb}$ in main paper).

7 Baselines and evaluation metrics

Here, we provide a more detailed description of the video captioning baselines used in our paper.

- MP-LSTM [11]. As an elementary baseline, we compare against an initial work that adopted a Recurrent Neural Network (RNN) for video captioning. This work combined a 2D-CNN, which is used for single image feature extraction, with a RNN, which is for caption decoding.
- TA [12]. This work augmented the previous CNN-LSTM combination model by: (i) introducing a temporal attention mechanism to exploit global temporal structure and (ii) substituting the visual feature extractor with a 3D-CNN, pre-trained under action recognition dataset.
- Transformer [10]. This method is one of the pioneer works that tried a transformer framework for video captioning. It still followed the encoder-decoder structure, but implemented with stacked self-attention units.
- UniVL [8]. UniVL is one of the top performers that build on transformers in recent years. It makes full use of the advantages of large-scale dataset pre-training (i.e., optimized through five pre-training tasks on HowTo100M [9]), and multi-modal feature cross encoding (i.e., vision and language). It achieved

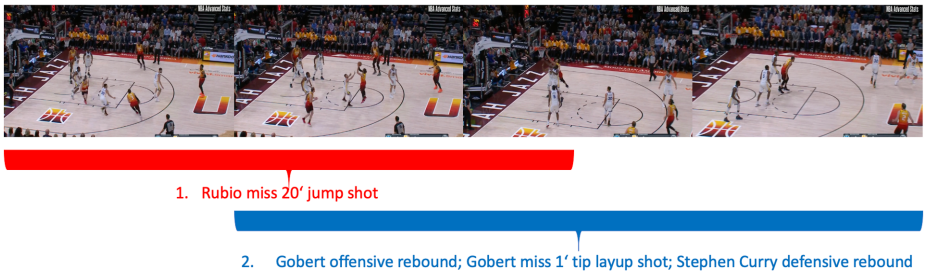

Fig. 5: The example shows two consecutive video-text pairs where the videos overlap. The color bars below video frames show the duration of each video. We can see that four events appear in the second video while the second text misses the first event, i.e., Rubio miss 20' jump shot, which is contained in the preceding video-text pair. We will target this problem and solve it for the next version of our dataset.

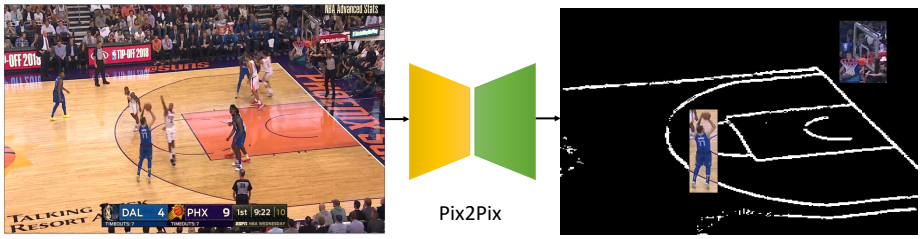

Fig. 6: Demonstration of courtline segmentation result produced by [13] (binary map in right sub-figure) and the positional-aware feature (right sub-figure) used in our approach.

SoTA results on multiple benchmarks: captioning, retrieval, action localization and multi-modality classificiation.

8 Analysis of feature impact

From Table 3 in main paper, it is seen that all features are individually effective, e.g., leading to notable improvements over TimeSformer feature alone. However, we observe that a plateau occurs when using more features with TimeSformer feature, i.e., T+BAL+BAS+PB+PA. This observation is more conspicuous on Meteor metric where adding PB+PA feature does not bring any further improvement. We consider three possible explanations for this phenomenon: (1) The combination of features has saturated the dataset. (2) In its formal definition [4], the Meteor score is defined as $F_{mean} \times (1 - \text{Penalty})$. The second term, i.e., $(1 - \text{Penalty}) \in [0, 5, 1)$, therefore discounts the performance improvements. (3) The Meteor score increases most when evaluating captions that have identical

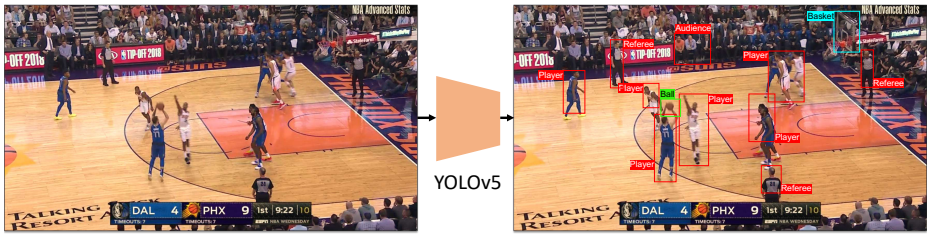

Fig. 7: Demonstration of player, ball and basket detection results used in our approach. Red label denotes players (only for illustrative purpose, we do not further differentiate players, referees and audiences in our approach), green label denotes the ball and cyan label the basket.

stem (e.g., missing vs. missed) as well as synonym (e.g., passing vs hand-over) with the ground truth. The presence of such ambiguities in NSVA is very minor.

9 Evaluation metrics for action and identity

Given a video clip in NSVA, the ground truth for action recognition as well as player identification is a sequence of data entries, where both the **order** and **correctness** matters. Therefore, we follow previous work [6,5] and evaluate the performance of our approach using three increasingly strict metrics. (1) mean Intersection over Union (mIoU) treats the predicted and ground-truth action sequences as sets, and measures the overlap between these sets. mIoU is agnostic to the order of actions and only indicates whether the model captures the correct set of steps needed to complete the plan. (2) mean Accuracy (mAcc) performs element-wise comparisons between the predicted and ground-truth action sequences, thereby considering the order of the actions or player names as well. (3) Success Rate (SR) considers a recognition successful only if it exactly matches the ground truth.

10 Full action space

Here, we provide the full list of action names in Table 1.

Table 1: Action name list of NSVA dataset

Action Name	Action ID
Block	1
Ejection-Other	2

Continued on next page

Table 1: Action name list of NSVA dataset (Continued)

Action Name	Action ID
Foul-Away-From-Play	3
Foul-Clear-Path	4
Foul-Defense-3-Second	5
Foul-Delay-Technical	6
Foul-Double-Personal	7
Foul-Double-Technical	8
Foul-Excess-Timeout-Technical	9
Foul-Flagrant-Type-1	10
Foul-Flagrant-Type-2	11
Foul-Hanging-Technical	12
Foul-Loose-Ball	13
Foul-Non-Unsportsmanlike-Technical	14
Foul-Offensive	15
Foul-Offensive-Charge	16
Foul-Personal	17
Foul-Personal-Take	18
Foul-Shooting	19
Foul-Technical	20
Foul-Too-Many-Players-Technical	21
Free-Throw-Free-Throw-1-of-1	22
Free-Throw-Free-Throw-1-of-2	23
Free-Throw-Free-Throw-1-of-3	24
Free-Throw-Free-Throw-2-of-2	25
Free-Throw-Free-Throw-2-of-3	26
Free-Throw-Free-Throw-3-of-3	27
Free-Throw-Free-Throw-Clear-Path-1-of-2	28
Free-Throw-Free-Throw-Clear-Path-2-of-2	29
Free-Throw-Free-Throw-Flagrant-1-of-1	30
Free-Throw-Free-Throw-Flagrant-1-of-2	31
Free-Throw-Free-Throw-Flagrant-1-of-3	32
Free-Throw-Free-Throw-Flagrant-2-of-2	33
Free-Throw-Free-Throw-Flagrant-2-of-3	34
Free-Throw-Free-Throw-Flagrant-3-of-3	35
Free-Throw-Free-Throw-Technical	36
Free-Throw-Free-Throw-Technical-1-of-2	37
Free-Throw-Free-Throw-Technical-2-of-2	38
Instant-Replay-Altercation-Ruling	39
Instant-Replay-Overturn-Ruling	40
Instant-Replay-Support-Ruling	41
Jump-Ball	42
Made-Shot-Alley-Oop-Dunk-Shot	43
Made-Shot-Alley-Oop-Layup-shot	44
Made-Shot-Cutting-Dunk-Shot	45
Made-Shot-Cutting-Finger-Roll-Layup-Shot	46
Made-Shot-Cutting-Layup-Shot	47
Made-Shot-Driving-Bank-Hook-Shot	48
Made-Shot-Driving-Dunk-Shot	49
Made-Shot-Driving-Finger-Roll-Layup-Shot	50
Made-Shot-Driving-Floating-Bank-Jump-Shot	51

Continued on next page

Table 1: Action name list of NSVA dataset (Continued)

Action Name	Action ID
Made-Shot-Driving-Floating-Jump-Shot	52
Made-Shot-Driving-Hook-Shot	53
Made-Shot-Driving-Layup-Shot	54
Made-Shot-Driving-Reverse-Dunk-Shot	55
Made-Shot-Driving-Reverse-Layup-Shot	56
Made-Shot-Dunk-Shot	57
Made-Shot-Fadeaway-Jump-Shot	58
Made-Shot-Finger-Roll-Layup-Shot	59
Made-Shot-Floating-Jump-shot	60
Made-Shot-Hook-Bank-Shot	61
Made-Shot-Hook-Shot	62
Made-Shot-Jump-Bank-Shot	63
Made-Shot-Jump-Shot	64
Made-Shot-Layup-Shot	65
Made-Shot-Pullup-Jump-shot	66
Made-Shot-Putback-Dunk-Shot	67
Made-Shot-Putback-Layup-Shot	68
Made-Shot-Reverse-Dunk-Shot	69
Made-Shot-Reverse-Layup-Shot	70
Made-Shot-Running-Alley-Oop-Dunk-Shot	71
Made-Shot-Running-Alley-Oop-Layup-Shot	72
Made-Shot-Running-Dunk-Shot	73
Made-Shot-Running-Finger-Roll-Layup-Shot	74
Made-Shot-Running-Jump-Shot	75
Made-Shot-Running-Layup-Shot	76
Made-Shot-Running-Pull-Up-Jump-Shot	77
Made-Shot-Running-Reverse-Dunk-Shot	78
Made-Shot-Running-Reverse-Layup-Shot	79
Made-Shot-Step-Back-Bank-Jump-Shot	80
Made-Shot-Step-Back-Jump-shot	81
Made-Shot-Tip-Dunk-Shot	82
Made-Shot-Tip-Layup-Shot	83
Made-Shot-Turnaround-Bank-Hook-Shot	84
Made-Shot-Turnaround-Fadeaway-Bank-Jump-Shot	85
Made-Shot-Turnaround-Fadeaway-shot	86
Made-Shot-Turnaround-Hook-Shot	87
Made-Shot-Turnaround-Jump-Shot	88
Missed-Shot-Alley-Oop-Dunk-Shot	89
Missed-Shot-Alley-Oop-Layup-shot	90
Missed-Shot-Cutting-Dunk-Shot	91
Missed-Shot-Cutting-Finger-Roll-Layup-Shot	92
Missed-Shot-Cutting-Layup-Shot	93
Missed-Shot-Driving-Bank-Hook-Shot	94
Missed-Shot-Driving-Dunk-Shot	95
Missed-Shot-Driving-Finger-Roll-Layup-Shot	96
Missed-Shot-Driving-Floating-Bank-Jump-Shot	97
Missed-Shot-Driving-Floating-Jump-Shot	99
Missed-Shot-Driving-Hook-Shot	100
Missed-Shot-Driving-Layup-Shot	101

Continued on next page

Table 1: Action name list of NSVA dataset (Continued)

Action Name	Action ID
Missed-Shot-Driving-Reverse-Dunk-Shot	102
Missed-Shot-Driving-Reverse-Layup-Shot	103
Missed-Shot-Dunk-Shot	104
Missed-Shot-Fadeaway-Jump-Shot	105
Missed-Shot-Finger-Roll-Layup-Shot	106
Missed-Shot-Floating-Jump-shot	107
Missed-Shot-Hook-Bank-Shot	108
Missed-Shot-Hook-Shot	109
Missed-Shot-Jump-Bank-Shot	110
Missed-Shot-Jump-Shot	111
Missed-Shot-Layup-Shot	112
Missed-Shot-Pullup-Jump-shot	113
Missed-Shot-Putback-Dunk-Shot	114
Missed-Shot-Putback-Layup-Shot	115
Missed-Shot-Reverse-Dunk-Shot	116
Missed-Shot-Reverse-Layup-Shot	117
Missed-Shot-Running-Alley-Oop-Dunk-Shot	118
Missed-Shot-Running-Alley-Oop-Layup-Shot	119
Missed-Shot-Running-Dunk-Shot	120
Missed-Shot-Running-Finger-Roll-Layup-Shot	121
Missed-Shot-Running-Jump-Shot	122
Missed-Shot-Running-Layup-Shot	123
Missed-Shot-Running-Pull-Up-Jump-Shot	124
Missed-Shot-Running-Reverse-Dunk-Shot	125
Missed-Shot-Running-Reverse-Layup-Shot	126
Missed-Shot-Step-Back-Bank-Jump-Shot	127
Missed-Shot-Step-Back-Jump-shot	128
Missed-Shot-Tip-Dunk-Shot	129
Missed-Shot-Tip-Layup-Shot	130
Missed-Shot-Turnaround-Bank-Hook-Shot	131
Missed-Shot-Turnaround-Fadeaway-Bank-Jump-Shot	132
Missed-Shot-Turnaround-Fadeaway-shot	133
Missed-Shot-Turnaround-Hook-Shot	134
Missed-Shot-Turnaround-Jump-Shot	135
Rebound-Normal-Rebound	136
Rebound-Unknown	137
Steal	138
Substitution	139
Timeout-Regular	140
Turnover-3-Second-Violation	141
Turnover-5-Second-Violation	142
Turnover-8-Second-Violation	143
Turnover-Backcourt-Turnover	144
Turnover-Bad-Pass	145
Turnover-Discontinue-Dribble	146
Turnover-Double-Dribble	147
Turnover-Excess-Timeout-Turnover	148
Turnover-Illegal-Assist-Turnover	149
Turnover-Illegal-Screen-Turnover	150

Continued on next page

Table 1: Action name list of NSVA dataset (Continued)

Action Name	Action ID
Turnover-Inbound-Turnover	151
Turnover-Jump-Ball-Violation	152
Turnover-Kicked-Ball-Violation	153
Turnover-Lane-Violation	154
Turnover-Lost-Ball	155
Turnover-No-Turnover	156
Turnover-Offensive-Foul-Turnover	157
Turnover-Offensive-Goaltending	158
Turnover-Out-of-Bounds—Bad-Pass-Turnover	159
Turnover-Out-of-Bounds-Lost-Ball-Turnover	160
Turnover-Palming-Turnover	161
Turnover-Punched-Ball-Turnover	162
Turnover-Shot-Clock-Turnover	163
Turnover-Step-Out-of-Bounds-Turnover	164
Turnover-Traveling	165
Violation-Defensive-Goaltending	166
Violation-Delay-Of-Game	167
Violation-Double-Lane	168
Violation-Jump-Ball	169
Violation-Lane	170
Period-Start	171
Period-End	172

References

1. NBA website box score. https://www.nba.com/game/mil-vs-lal-0021900939/box-score 4
2. NBA website game chart. https://www.nba.com/game/mil-vs-lal-0021900939/game-chart 4
3. NBA website game summary. https://www.nba.com/game/mil-vs-lal-0021900939 4
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of ACL (2005) 7
5. Bi, J., Luo, J., Xu, C.: Procedure planning in instructional videos via contextual modeling and model-based policy learning. In: Proceedings of ICCV (2021) 8
6. Chang, C.Y., Huang, D.A., Xu, D., Adeli, E., Fei-Fei, L., Niebles, J.C.: Procedure planning in instructional videos. In: Proceedings of ECCV (2020) 8
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of CVPR (2017) 5
8. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Chen, X., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. CoRR **abs/2002.06353** (2020) 6
9. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of CVPR (2019) 6

10. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of ACL (2018) 6
11. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Proceedings of NAACL (2015) 6
12. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of ICCV (2015) 6
13. Zhu, L., Rematas, K., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Reconstructing NBA players. In: Proceedings of ECCV (2020) 5, 7