Fine-grained Video Captioning for Sports Narrative
Supplementary Material

Huanyu Yu*, Shuo Cheng*, Bingbing Ni*†, Minsi Wang, Jian Zhang, Xiaokang Yang
Shanghai Institute for Advanced Communication and Data Science,
Shanghai Key Laboratory of Digital Media Processing and Transmission,
Shanghai Jiao Tong University
yiranyhy@163.com, acccheng94@gmail.com, nibingbing@sjtu.edu.cn,
mswang1994@gmail.com, stevenash0822@sjtu.edu.cn, xkyang@sjtu.edu.cn

Appendix Overview

In this document we include additional material related to dataset and evaluation. In **Appendix I** we display the rich and accurate annotations about the dataset, and in **Appendix II** we compare our evaluation metric with other main stream conventional metrics in details.

Appendix I: More Details About FSN Dataset

FSN dataset is a fine-grained sports video captioning dataset. Due to the professionalism of sports videos, the captioning labels are accomplished by a professional team which consists of 20 persons who are equipped with extensive basketball knowledge and experience instead of the Amazon Mechanical Turk to ensure the the quality of the dataset in terms of the description accuracy of action details. Each video is annotated by at least three persons, and the final annotation is chosen by their agreements. In addition, we densely annotate each player and the ball in every frame with different tags (where 0 represent the background, the players are tagged with 1-10, and the ball is tagged with 11, respectively). It takes 1 month to accomplish the annotation procedure. In Figure 3, we visualize some randomly selected sequences in our dataset.

Because of the high quality of the video frames and the professional annotations, our dataset can also be used for other vision tasks (*e.g.*, video segmentation, tracking, action recognition), or for pre-training other deep learning models. The full dataset will be released to encourage related researches.

Appendix II: Comparison of Evaluations

Fine-grained Captioning evaluation (FCE) metric is designed to pay more attention to the accuracy of the verbs in the sentence since the verbs play a crucial role in sports video description. Verbs account for a large proportion in FSN dataset as is visualized in Figure 1 and Figure 2. The similar distribution of the verbs' number in the reference sentences and the candidate sentence in these two figure also indicate that the output of the fine-grained video captioning model is reasonable.

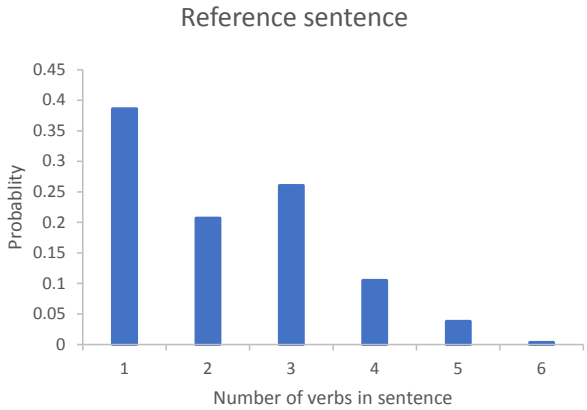


Figure 1. The probability distribution of the verbs' number in the reference sentences in FSN dataset.

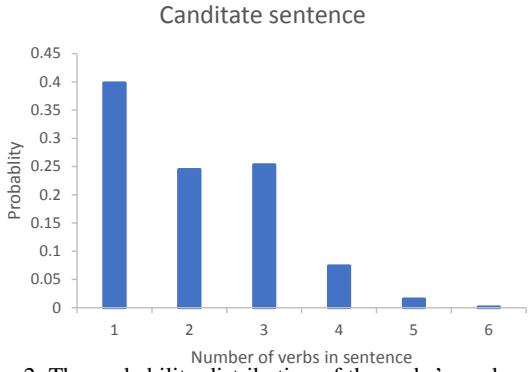
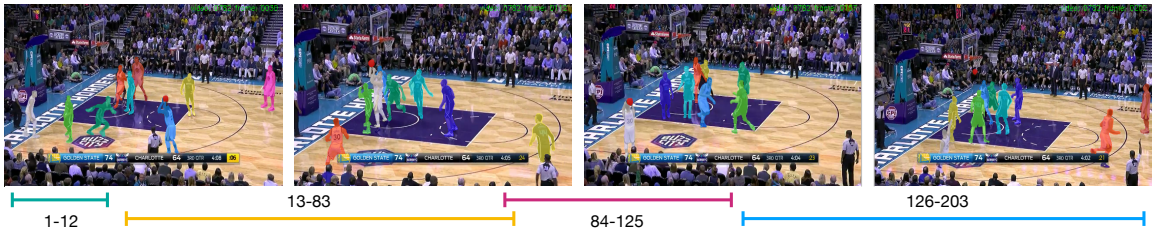
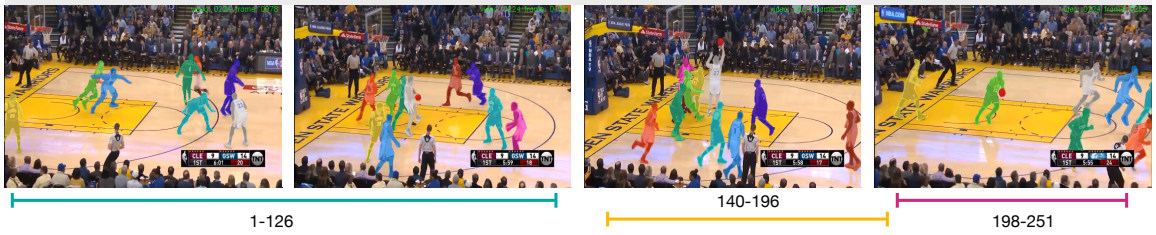


Figure 2. The probability distribution of the verbs' number in the candidate sentences in FSN dataset.

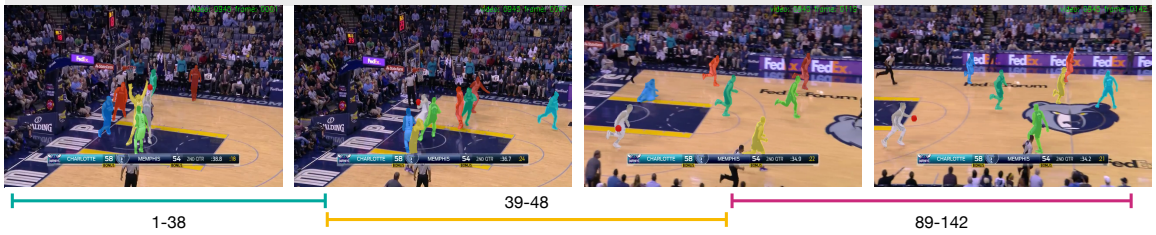
*Authors contributed equally to this work.
†Corresponding Author.



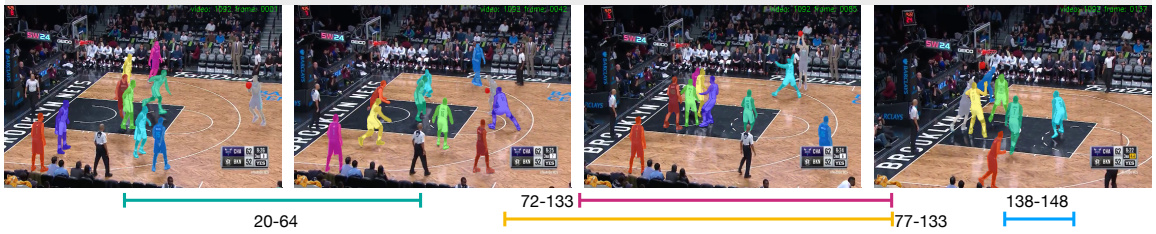
Reference: A person passes the ball to his teammate. The teammate makes a three-point shot but does not score. The teammate gets the rebound and passes the ball to another teammate. The ball handler make a three-point shot and scores.



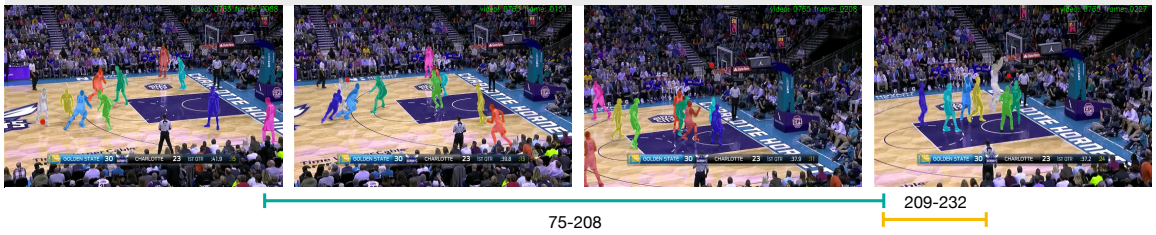
Reference: A man dribbles the ball across the court and passes the ball to his teammate. The ball handler shoots the ball and scores a basket. A defender gets the ball and passes the ball to his teammate.



Reference: A person drives to the hoop but is blocked by the opponent. The defender gets the rebound and passes the ball to his teammate. The ball hander dribbles the ball forward.



Reference: A person dribbles the ball forward and passes the ball to his teammate. The teammate makes a three-point shot and scores. The defender raises his hands but fails to block him. The defender gets the ball.



Reference: A person bypasses the opponent's defense and makes a three-point shot but does not score. The teammate gets the rebound and makes a slam dunk and scores.

Figure 3. Sample sequences from our dataset. Each player and the ball has accurate pixel-level annotation, the figures below/above the line indicate the duration of each captioning event, with caption highlighted in the same color. Best viewed in colors.

Couple 1	**Candidate Sentence(C) and Reference Sentence(R)**	**METEOR**	**FCE**
Better Caption	C: A person dribbles the ball forward and drives to the hoop and scores R: A person dribbles forward and drives to the hoop and shoots and scores.	0.45	0.45
Worse Caption	C: A man pretends to shoot the ball but passes the ball to his teammate R: A person passes the ball to his teammate.	0.45	0.26
Couple 2	**Candidate Sentence(C) and Reference Sentence(R)**	**METEOR**	**FCE**
Better Caption	C: The teammate makes a three-point shot and scores. R: The ball handler makes a three-point shot and scores.	0.41	0.41
Worse Caption	C: A person bypasses the opponent's defense and makes a three-point shot and scores. R: A person gets the ball and makes a three-point shot and scores.	0.41	0.32
Couple 3	**Candidate Sentence(C) and Reference Sentence(R)**	**METEOR**	**FCE**
Better Caption	C: A person gets the ball and makes a three-point shot and scores. R: The teammate makes a shot and scores.	0.40	0.40
Worse Caption	C: A man bypasses the defender and makes a jump shot but does not score. R: A person gets the ball and makes a jump shot but does not score.	0.40	0.36

Table 1. Good caption and bad caption with the same METEOR scores get different FCE scores. All the well described sentences get the higher FCE scores than the badly described sentences.

Metric	Correlation
Precision	0.457
Recall	0.526
METEOR	0.579
FCE	**0.598**

Table 2. Correlations between human evaluations and precision, recall, METEOR and FCE Scores.

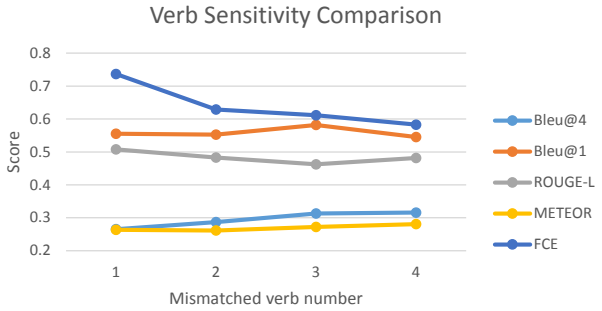


Figure 4. Illustration of the relationship between metric score and the number of mismatch verb.

As noted in the main paper, the FCE metric can evaluate the verb accuracy and the order of the motion appearing in captioning sentences.To evaluate this new metric, we compare FCE metric with other metrics in three aspects.

First we consider the correlation with human assessment since people always have a good intuition about the quality of the candidate sentences. We ask 20 persons to assess the 500 candidate sentences with the according reference sentences. Each score ranges from one to five (with one being the poorest grade and five being the highest) and each sentence is evaluated by at least two separate human judges. The final score of the sentence is computed by averaging the whole scores. The human judges are asked to assess the candidate sentence from both fluency and adequacy especially the motion accuracy. Then we compute the sentence by sentence correlation between human assessments and several metrics including FCE, METEOR, Recall and Precision. Table 2 shows the sentence level correlation between human judges and these metrics over the FSN dataset. By taking verb accuracy into account, our FCE metric get further improvement in correlation.

Second we conduct a comparison between FCE and traditional METEOR metrics. We choose sufficient pairs of better-motion-described candidate sentence and worse-motion-describe candidate sentence, which share the similar scores by METEOR, and evaluate these pairs of well-poor sentences by FCE metric. We observe that the score of the worse-described one is significantly lower than the better-described one by FCE metric. Thus, the rationality of this FCE metric is verified. Table 1 displays some examples of the couples.

For analyzing the verb error tolerance about each evaluation metric, we visualize the relationship between the metric scores and the number of mismatch verb terms, see Figure 4 for more details.

We find our evaluation metric is more sensitive about verb errors. When the number of mismatch verb terms increase, our evaluation score drops significantly compared with other metrics, further demonstrates the effectiveness and rationale of the FCE metric for fine-grained video captioning tasks. On the other hand, we observe the fact that other conventional metrics almost remain stable with variation of number of mismatch verb terms, which points out that conventional evaluation metrics are not sufficient for evaluating fine-grained sports captioning tasks (as they only evaluate the semantic level similarity between candidate and reference sentence, but not focus on verbs, which are more crucial for judging the caption quality). Our proposed metric manages to compensate this shortage.