

# Player-Centric Multimodal Prompt Generation for Large Language Model Based Identity-Aware Basketball Video Captioning

## Supplementary Material

### A. Identity-aware Basketball Video Captioning Dataset

With the improvement of people’s living standards, the vigorous development of sports undertakings, and the growing enthusiasm of the public for sports, the demand for sports video understanding is also expanding. In sports video captioning, there are typically two types of descriptions: 1) commentary, which targets TV audiences and tends to be long and detailed; 2) live text broadcasting, which targets online users and is more concise. Our work focuses on the latter, where the captions are brief but include key events with corresponding players. Moreover, with the fast pace of life and busy work schedules, people may not have the time to watch sports videos. Compared to videos, which require people to spend a certain amount of time to understand the game situation, text live broadcasts can provide people with concise and clear text, allowing them to quickly grasp the information of the videos. Therefore, for basketball live text broadcast, we aim to construct a dataset whose text provides concise event descriptions, including actions and participants, helping audiences quickly grasp key content and ensuring efficient updates and summaries for sports fans.

The proposed NBA-Identity dataset is designed for basketball live text broadcasting. Each video clip of an event is annotated with a corresponding description. Unlike traditional video captioning datasets that summarize video content with broad overviews, this dataset focuses on providing descriptions with specific player identities and actions to describe visual content. This section introduces the following aspects of the NBA-Identity dataset: 1) dataset collection process, 2) statistical analysis, and 3) dataset versatility.

#### A.1. Dataset Collection

We collect text data from 40 games through the professional basketball data platform and obtain corresponding video data from a basketball live streaming platform. Using the methodology of VC-NBA-2022 [10], the text and video data are aligned through Tesseract-OCR [6]. The event types we used are sourced from professional basketball live broadcast data. A few low-frequency event types, such as “jump ball” (about 1/350 of all events per game) and “violation” (about 1/350 of all events per game), are excluded. From these 40 games, we extract a total of 9,726 video clips, covering 321 player identities and 9 major types of events: “block”, “foul”, “defensive rebound”, “offensive rebound”, “turnover”, “two-point (2-pt) shot”,


VideoID: 100001

Captions: Offensive rebound by C. LeVert
Bbox: [[701, 349, 787, 500], [768, 348, 846, 499], [841, 345, 902, 510], [859, 358, 981, 517], [939, 381, 996, 524], [963, 410, 1028, 542], [953, 400, 1047, 542], [949, 395, 1059, 518], [933, 389, 1026, 526], [920, 372, 1013, 559], [864, 363, 990, 564], [850, 352, 924, 559]]
Action: Offensive rebound
Player: C. LeVert
Start and end time: [4, 7.6]

Figure 1. Data samples from the proposed dataset. Each video clip is annotated by video\_id, caption, bounding box, action type and player names.

“three-point (3-pt) shot”, “layup”, and “assist”. The nine major event types mentioned in the paper are broad categories, but for example, “foul” includes “personal foul”, “personal take foul”, “shooting foul”, “offensive foul”, and “loose ball foul”. Similarly, “turnover” includes “lost ball; steal by [player]”, “bad pass; steal by [player]”, “lost ball”, “bad pass”, “offensive foul”, “traveling”, “step out of bounds”, “discontinued dribble”, and “out of bounds lost ball”. “Shot” events are further divided into “make shot” and “miss shot”. The dataset also contains events with limited samples, such as “free throw” and “drawn”. Our dataset actually covers a total of 26 fine-grained event types in basketball. Notably, the VC-NBA-2022 dataset discards shooting distance (e.g., “from 20 ft”) and combines missed shots with rebounds into a single event (e.g., “L. Shamet misses the 3pt jump shot and E. Gordon gets the defensive rebound”). In contrast, we retain the shooting distance to enhance the dataset’s complexity and authenticity. The missed shots and rebounds are not merged because it could lead to increased video duration, and the rapid transitions of scenes would make the identification of these events easier. Furthermore, we annotate the coordinates of the key players for each video clip. Key players refer to the players included in the description of the video clip. The dataset sample is shown in Fig. 1. NBA-Identity is split by game instances, with 35 games (8,667 clips) for training and 5 games (1,059 clips) for testing.

#### A.2. Dataset Statistics

Our dataset is the largest identity-aware video captioning dataset in sports domain, both in terms of the number of

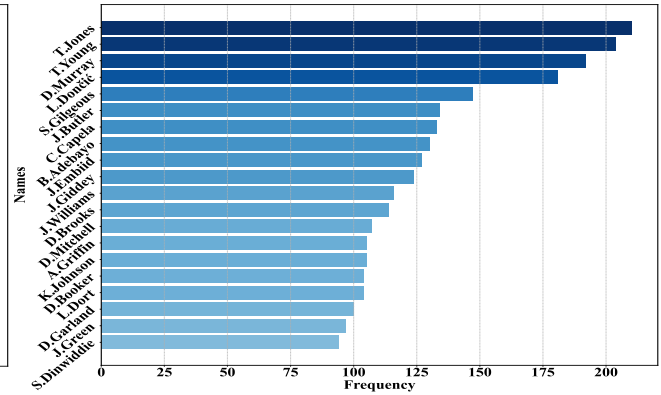
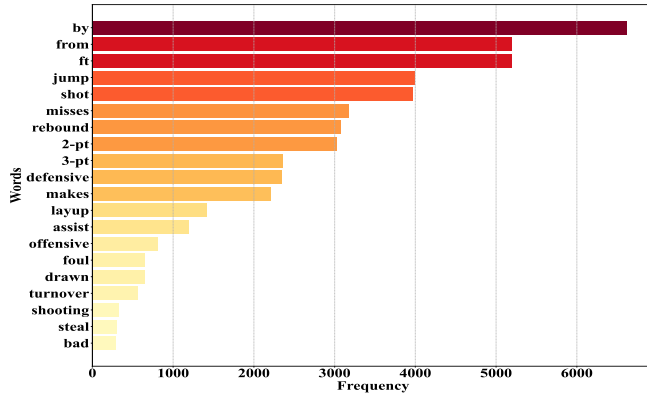


Figure 2. Illustrations of NBA-Identity dataset statistics.

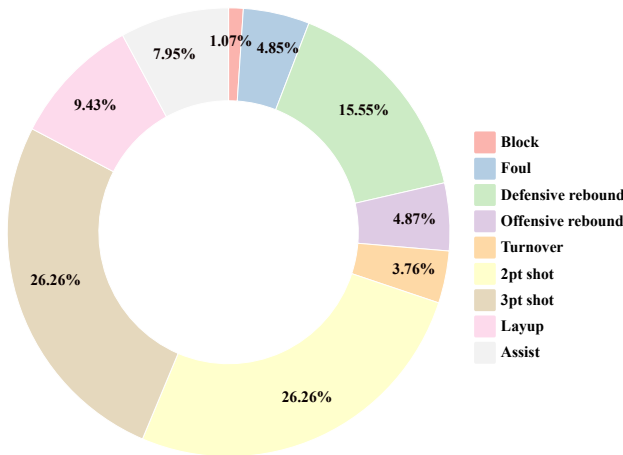


Figure 3. The distribution of basketball action categories contained in NBA-Identity dataset.

videos and descriptions, with a particular emphasis on annotating player identities. This provides researchers with a richer data resource to support in-depth analysis and understanding of basketball events. The distribution of the top 20 most frequent words in the dataset is shown in Fig. 2 (a), illustrating common descriptive vocabulary. Fig. 2 (b) presents the frequency of player identities, with the highest occurrence being that of T.Jones. These statistics offer insights into the trends in descriptions and player involvement, providing valuable references for subsequent modeling efforts. Additionally, Fig. 3 illustrates the distribution of different event types in the dataset. Shooting and rebounding events occur most frequently, while blocking events are the least common. This distribution aligns with real-game scenarios, reflecting the typical frequency of various events in basketball games. We also provide word-cloud-based statistics in Fig. 4 to reveal the relative amount of different words. It shows that the top-5 subjects in NBA-Identity are

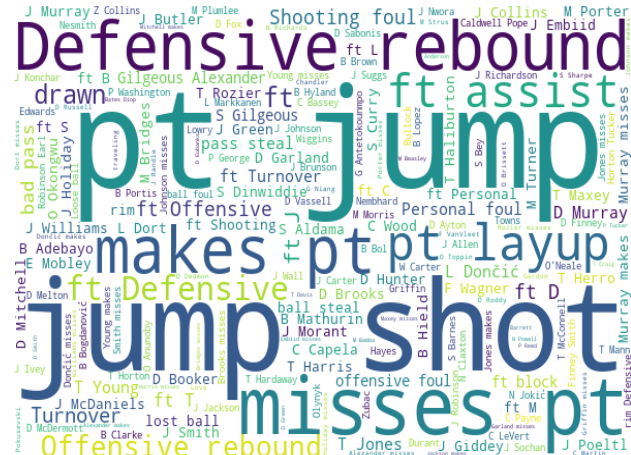


Figure 4. Word cloud of NBA-Identity dataset. The bigger the font, the more percentage it occupies.

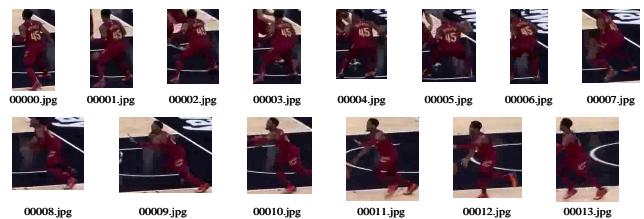


Figure 5. Example of player sequence.

“jump”, “shot”, “pt”, “misses” and “defensive”, followed by “rebound”, “makes”, “layup” and “defensive”.

### A.3. Dataset Versatility

This dataset demonstrates significant versatility and development potential, offering a valuable resource for future research. In addition to providing detailed descriptions for each video clip, the dataset includes annotations for ac-





Figure 6. Example tracking results of SportsMOT on the test set of NBA-Identity. Each row shows the results of sampled frames in chronological order of a video clip. Bounding boxes and players are marked in the images. Bounding boxes with different colors represent different players.

tion types, player coordinates, and the temporal boundaries of events. These comprehensive annotations extend the dataset’s application range, supporting not only video captioning tasks but also enabling research in group activity recognition [9], player identity recognition [7] and temporal action detection [3]. Furthermore, the multidimensional annotation details offer researchers opportunities to explore correlations between video data and event characteristics, laying a strong foundation for in-depth research in the field of sports analysis.

## B. Evaluation Metrics

We employ typical captioning metrics (*e.g.*, BLEU (B) [5], Rouge-L (R) [4], METEOR (M) [1] and CIDEr (C) [8]) to

evaluate the performance of LLM-IAVC.

(1) BLEU is one of the most widely used metrics in machine translation and captioning. It evaluates the quality of generated captions by calculating the precision of  $n$ -gram matches between the generated and reference captions, typically using BLEU-4 (4-gram) as the standard.

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N \log p_n \right), \quad (1)$$

where  $p_n$  is the  $n$ -gram precision, calculated as the number of matched  $n$ -grams divided by the total number of  $n$ -grams in the generated caption. BP denotes the Brevity Penalty, which penalizes overly short generated captions:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}, \quad (2)$$

where  $c$  denotes the length of the generated caption and  $r$  denotes the length of the shortest reference caption.  $w_n$  is the weight for  $n$ -grams, typically set to  $w_n = \frac{1}{N}$  (e.g.  $N = 4$  for BLEU-4).

(2) Rouge-L evaluates the quality of generated captions by calculating the longest common subsequence (LCS) between the generated and reference captions, with a focus on recall (R).

$$R_{\text{LCS}} = \frac{\text{LCS}(\mathbf{X}, \mathbf{Y})}{\text{len}(\mathbf{Y})}, \quad (3)$$

$$P_{\text{LCS}} = \frac{\text{LCS}(\mathbf{X}, \mathbf{Y})}{\text{len}(\mathbf{X})}, \quad (4)$$

$$\text{Rouge-L} = \frac{(1 + \beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}, \quad (5)$$

where  $\mathbf{X}$  and  $\mathbf{Y}$  denote the generated caption and the reference caption, respectively.  $\text{LCS}(\mathbf{X}, \mathbf{Y})$  denotes the length of the longest common subsequence between  $\mathbf{X}$  and  $\mathbf{Y}$ . And  $\beta$  is the weight parameter, typically set to 1 (balancing recall and precision).  $P$  is the precision score.

(3) METEOR combines precision (P), recall (R), and word order in its evaluation, closer to human evaluation. It incorporates synonym matching and stemming to improve robustness.

$$\text{METEOR} = (1 - \gamma \cdot \text{Penalty}) \cdot \frac{P \cdot R}{\alpha P + (1 - \alpha) R}, \quad (6)$$

where  $P$  denotes the precision score, calculated as the number of matched words divided by the total number of words in the generated caption.  $R$  denotes the recall score, calculated as the number of matched words divided by the total number of words in the reference caption.  $\alpha$  is the weight parameter for balancing precision and recall, typically set to 0.9.  $\gamma$  is the the penalty weight, typically set to 0.5. And Penalty denotes the word order penalty, calculated based on the difference in word order between the generated and reference captions.

(4) CIDEr is specially designed for captioning task evaluation. It evaluates the semantic consistency of generated captions by computing the similarity between the generated and reference captions using TF-IDF weighting.

$$\text{CIDEr}(\mathbf{C}, \mathbf{S}) = \frac{1}{m} \sum_{j=1}^m \frac{g(\mathbf{C}) \cdot g(\mathbf{s}_j)}{\|g(\mathbf{C})\| \cdot \|g(\mathbf{s}_j)\|}, \quad (7)$$

where  $\mathbf{C}$  denotes the generated caption.  $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m\}$  denotes the set of reference captions.  $g(\cdot)$  denotes the TF-IDF vector representation of a caption. The cosine similarity between the generated caption and each reference caption is calculated and averaged.

Model	Down_dim	CIDEr	BLEU-4
LLM-IAVC (Llama3.2-3B)	128	99.6	15.8
	256	100.1	17.2
	512	105.3	18.8
	768	105.3	18.0

Table 1. Impact of dimension settings in down-projection matrix.

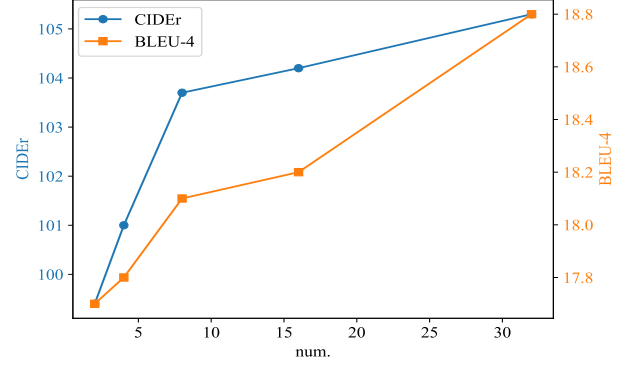


Figure 7. Ablation on the number of learnable vectors in VCLM.

### C. SportsMOT: Multi-object Tracker

At the training stage, player sequences are cropped from the video based on the bounding boxes provided in the dataset, as shown in Fig. 5. Subsequently, the player recognition module extracts visual features from these sequences for further processing. However, providing player bounding boxes directly during the inference stage is unreasonable and impractical. Therefore, at the inference stage, the multi-objective tracker SportsMOT [2] is utilized to extract player sequences of multiple players. It focuses on multi-player tracking and excludes audience members and referees. Our proposed player identification network extracts their visual features and obtain corresponding player names from player sequences. The multi-object tracking results are shown in Fig. 6.

### D. Additional Ablation Studies

**Effects of Different Dimensions of BSIM.** BSIM initially compresses the input features into a lower dimension, subsequently performs interaction operations, and ultimately reconstructs the dimensions back to their original size. This approach improves computational efficiency by reducing redundant information. It also ensures that the interaction emphasizes the most important features, which in turn enhances the model’s performance and generalization ability. We also conduct experiments comparing the BSIM module’s performance with different intermediate lower dimensions. As shown in Tab. 1, when the dimension is set to 512, the model achieves optimal performance, with CIDEr at 105.3 and BLEU-4 at 18.8. This design allows BSIM



to retain essential information while reducing unnecessary computational overhead, ultimately improving the overall efficiency and effectiveness of the model.

#### **Ablation on the number of learnable vectors in VCLM.**

We exploit how the number of learnable query vectors affects LLM-IAVC performance. As shown in Fig. 7, the model performs best with 32 vectors. Performance declines when the number is below 18, as fewer tokens capture less key video content. Conversely, using more than 32 vectors introduces redundant information. Ultimately, the number of query vectors for VCLM is set to 32.

## **E. Discussion**

**Dataset.** Compared to existing identity-aware sports video captioning datasets, our dataset contains the largest number of videos. Each visual scene varies with changes in lighting, player positions, and the appearance of actions. Given the high annotation cost, like other sports video captioning datasets, each video has only one description. These captions crawled from live text commentary websites primarily include different player names, actions, and distances. In addition, captions on our dataset aim to help audiences quickly understand match information (specifically which player performs what action). As a result, the caption does not need to prioritize diversity, creativity, or rich expressions. Each action is described using official and professional terminology.

**The generality of pipeline.** Our method is a highly generalizable framework that can be seamlessly applied to a wide range of sports, including but not limited to volleyball, table tennis, baseball, and soccer, without requiring any modifications to its core modules. However, it requires manually annotating player bounding boxes to pre-train the corresponding player identification network for each sport. This step is crucial for enabling the model to accurately identify and track players within the unique visual contexts and dynamics of each sport.

## **References**

- [1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 65–72, 2005. 3
- [2] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, et al. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9921–9931, 2023. 4
- [3] Kai Hu, Chaowen Shen, Tianyan Wang, et al. Overview of temporal action detection based on deep learning. *Artificial Intelligence Review*, 57(2):26, 2024. 3
- [4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3
- [5] Kishore Papineni, Salim Roukos, et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. 3
- [6] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, pages 629–633. IEEE, 2007. 1
- [7] Kanav Vats, Pascale Walters, Mehrnaz Fani, et al. Player tracking and identification in ice hockey. *Expert systems with applications*, 213:119250, 2023. 3
- [8] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 3
- [9] Lifang Wu, Meng Tian, Ye Xiang, et al. Learning label semantics for weakly supervised group activity recognition. *IEEE Transactions on Multimedia*, 2024. 3
- [10] Zeyu Xi, Ge Shi, Xuefen Li, et al. A simple yet effective knowledge guided method for entity-aware video captioning on a basketball benchmark. *Neurocomputing*, 619:129177, 2025. 1