

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

Zeyu Xi¹, Haoying Sun¹, Yaofei Wu¹, Junchi Yan², Haoran Zhang¹, Lifang Wu^{1*},
 Liang Wang³, Changwen Chen⁴

¹Beijing University of Technology; ²Shanghai Jiao Tong University; ³Chinese Academy of Sciences

⁴The Hong Kong Polytechnic University

*Corresponding author: lfwu@bjut.edu.cn

Abstract

Existing sports video captioning methods often focus on the action yet overlook player identities, limiting their applicability. Although some methods integrate extra information to generate identity-aware descriptions, the player identities are sometimes incorrect because the extra information is independent of the video content. This paper proposes a player-centric multimodal prompt generation network for identity-aware sports video captioning (LLM-IAVC), which focuses on recognizing player identities from a visual perspective. Specifically, an identity-related information extraction module (IRIEM) is designed to extract player-related multimodal embeddings. IRIEM includes a player identification network (PIN) for extracting visual features and player names, and a bidirectional semantic interaction module (BSIM) to link player features with video content for mutual enhancement. Additionally, a visual context learning module (VCLM) is designed to capture the key video context information. Finally, by integrating the outputs of the above modules as the multimodal prompt for the large language model (LLM), it facilitates the generation of descriptions with player identities. To support this work, we construct a new benchmark called NBA-Identity, a large identity-aware basketball video captioning dataset with 9,726 videos covering 9 major event types. The experimental results on NBA-Identity and VC-NBA-2022 demonstrate that our proposed model achieves advanced performance. Code and dataset are publicly available at <https://github.com/Zeyu1226-mt/LLM-IAVC>.

1. Introduction

Video captioning is a critical research topic in video understanding with the aims of generating precise descriptions to comprehend the visual contents of the given videos. It has attracted increasing attention due to a variety of its applica-

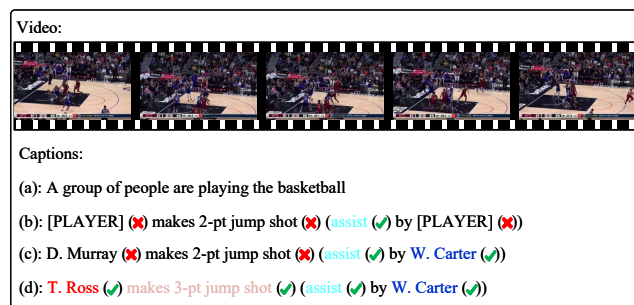


Figure 1. Caption comparison of current sports video captioning methods and our method. The different player names are marked red and blue. The actions are marked pink and cyan. The green checkmarks indicate correct name/action predictions, while the red crosses indicate incorrect ones.

tions including sports video understanding [24, 29, 33, 45, 47], video storytelling [19] and online video search [16, 25].

The traditional video captioning methods [21, 38, 39, 44, 50] achieve impressive results in open domains, which provide only a broad summary of the visual content, as shown in Fig. 1 (a). However, in some specific scenarios, such as sports live text broadcast, the players' identities and the actions are crucial, as audiences are particularly interested in these details. Such requirements cannot be satisfied by the general descriptions. Anonymous methods replace entities in descriptions with special tokens (e.g., [PLAYER] and [TEAM]) [24, 28, 33, 34] or numeric IDs (e.g., "PLAYER0961226") [45]. They can provide detailed actions and distinguish different entity types. However, their descriptions are still missing entity identities, specifically player names, as shown in Fig. 1 (b).

Recently, some methods [29, 47, 48] have introduced extra information, such as game news, player statistics, and the candidate player list, to generate descriptions with player identities. Due to the fact that the extra information is independent of the video clips, player identities in the descriptions are derived from the static extra information

rather than from the video clips themselves. These methods do not truly recognize player identities from the video clips, but rather make inferences based on extra information. The player identities may be incorrect, as shown in Fig. 1 (c).

The above methods fail to correctly describe players' identities, which has become a major backlog for sports video captioning. In fact, player identities are implied within the video clips, but most datasets [4, 6, 9, 30, 33, 34, 49, 54] do not include the entity names, which makes it impossible to generate the identity-aware descriptions. Although some existing sports video captioning datasets include player names [29, 47, 48], they lack the necessary annotations to facilitate the development of models with visual mining capabilities. For instance, player bounding boxes could help build models to directly recognize player identities in visual content. If we could re-organize video clips around individual players to obtain the player-centric video clip set, it is possible to develop a player identification network for identity-aware video captioning.

In this paper, we collect 40 NBA full games during 2022-2023, and organize a player-centric video clip set to link each player with all of his video clips. Based on this set, a player identification network (PIN) is developed to extract players' visual features and names from the given videos. A bidirectional semantic interaction module (BSIM) is designed to augment the player visual features by interacting with the video features. The above two modules form an identity-related information extraction module (IRIEM), which outputs the identity-related multimodal embeddings. A learnable approach is adapted by utilizing a designed visual context learning module (VCLM) to learn video context information, which is then concatenated with the identity-related multimodal embeddings as the prompt for the LLM to generate descriptions with player names. To support this, we propose an identity-aware basketball video captioning dataset NBA-Identity, which involves 9,726 videos, 321 players with bounding boxes, and 9 major kinds of events.

The main contributions of this paper are:

- We propose a player-centric multimodal prompt generation network for LLM-based identity-aware basketball video captioning (LLM-IAVC), which jointly utilizes player identity features and contextual video information to guide LLM in generating identity-aware descriptions.
- A large identity-aware basketball video captioning dataset NBA-Identity is constructed. This dataset contains the largest number of videos among identity-aware sports video captioning datasets. Its captions cover 9 major events in the basketball domain.
- Extensive experiments on our NBA-Identity and VC-NBA-2022 [47] datasets demonstrate that the proposed model achieves state-of-the-art performance without the need for additional knowledge assistance.

2. Related Works

Video Captioning. Video captioning aims to comprehend the video content and produce descriptions in natural language automatically. Recent researches [21, 38, 39, 44, 50] primarily emphasize generation methods based on sequence learning, where a visual encoder extracts relevant visual features from videos, and a decoder then generates descriptions in a sequential manner. Researchers have attempted to employ various visual encoders, including ResNet [13], Vision Transformer (ViT) [10], and C3D [12], to extract different 2D/3D video features. In this work, our LLM-IAVC employs the encoder-decoder framework and integrates the pre-trained model Timesformer [3] as the visual encoder to encode the spatiotemporal information in videos. Unlike most models that use LSTM [7] or Transformer [42] as the text decoder, we employ the LLM for the naturalness and accuracy of generated text.

Sports Video Captioning. It aims to recognize player identities and key event types, generating accurate descriptions in fluent and natural language to ensure alignment with the video content. Anonymous methods utilize special tokens (*e.g.*, [PLAYER] and [TEAM]) [24, 28, 33, 34] or entity types (*e.g.*, "Batter" in baseball) [15, 17] in descriptions to represent entities. These methods still cannot generate descriptions with specific entity names. Extra information based methods [29, 47, 48] generate descriptions that include player names by incorporating extra game-related information. This type of method guesses player identities from a pool of candidate information but does not genuinely recognize players' identities from a visual perspective. Different from the above methods, this paper focuses on extracting identity-related information directly from the video clips, which is helpful for preserving the consistency of the descriptions and the video clips.

Large Language Models. LLMs (*e.g.*, LLaMa2 [41] and GPT4 [1]) exhibit proficiency in a variety of language tasks given their powerful reasoning ability. Recent works [26, 36, 53] attempt to empower LLMs with visual understanding ability on downstream tasks like visual question-answering, video grounding, and video anomaly detection. With its impressive text generation capabilities, we employ various lightweight LLMs as the decoders, including GPT-2 [31], Qwen2.5 [40], and Llama3.2 [23], to generate high-quality descriptions for the sports video captioning task. Instead of manually designing a sentence as a prompt, we automatically generate a multimodal prompt that includes identity-related information and visual context, helping the LLM better understand the video content.

3. Methods

As shown in Fig. 2, LLM-IAVC mainly consists of three main components: 1) identity-related information extrac-

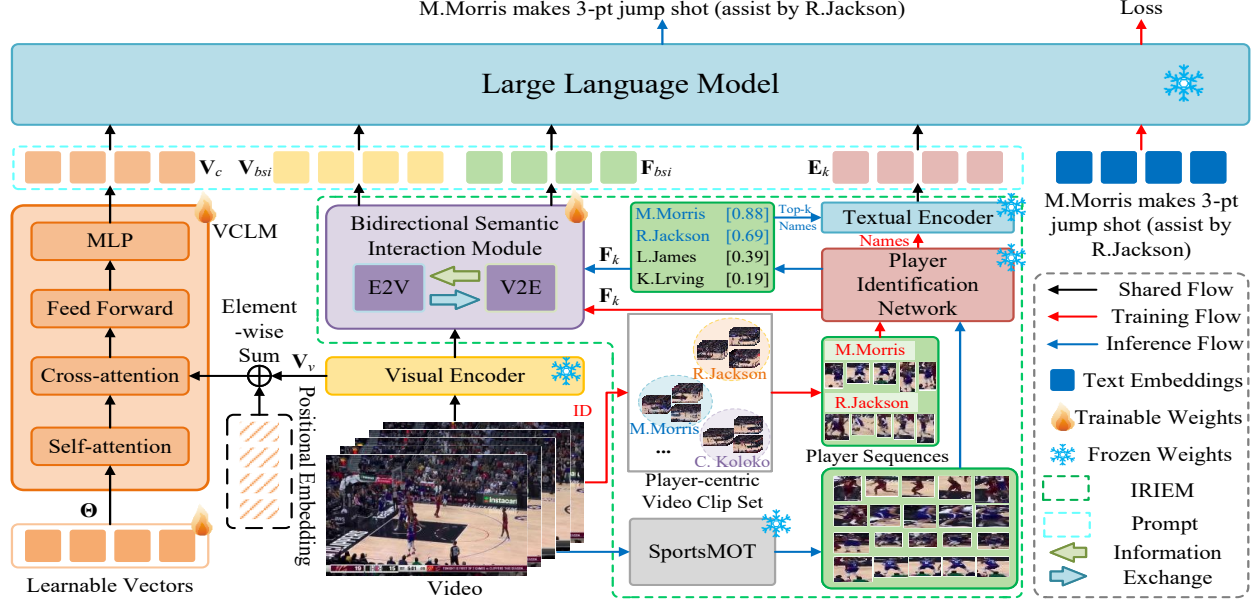


Figure 2. Illustration of the proposed identity-aware basketball video captioning model. During training, the player sequences corresponding to the input video are located from the player-centric video clip set based on the video ID, and their visual features are extracted via a pre-trained player identification network. During inference, a multi-object tracker extracts player sequences, and the player identification network extracts their visual features and identities. The top- k players with the highest confidence are selected for further processing.

tion module (IRIEM), 2) visual context learning module (VCLM), and 3) LLM-based decoder. The IRIEM includes the player identification network (PIN) and the bidirectional semantic interaction module (BSIM). The red line represents the training flow and the blue line represents the inference flow. The black flow represents the shared process for training and inference. In the following subsections, we will describe each component in detail.

3.1. Player Identification Network

Since player identities are inherently embedded in video clips, a straightforward approach is to develop a player identification module that directly recognizes player identities from visual content. Based on the player names in captions, each video clip is associated with specific players. Therefore, all training video clips from 40 full games are automatically re-organized as the player-centric video clip set. The player sequences could be obtained by their bounding boxes. The player identification network is developed to extract the identity-related visual features from the player sequences. Fig. 3 shows the training process of the player identification network, which consists of TimeSformer as the visual backbone and a fully connected layer as the classification head.

Let $P_B = \{p_1, p_2, \dots, p_n\}$ represent the player sequence, where p_n denotes the n -th the cropped bounding box. The player sequence feature $\mathbf{F}_P \in \mathbb{R}^{1 \times D_{time}}$ is extracted by TimeSformer [3]. Then $\mathbf{F}_P$ is fed into a classifi-

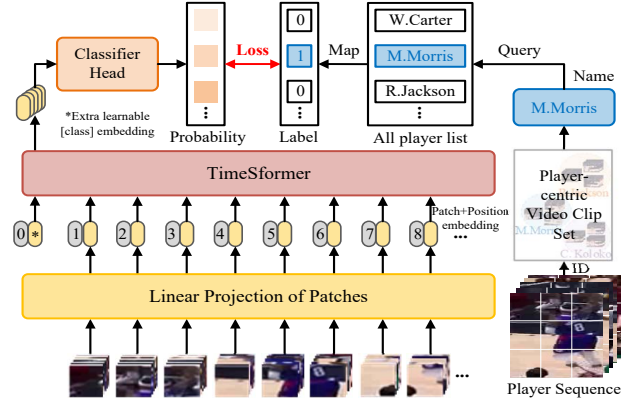


Figure 3. Training for the player identification network, including TimeSformer as the visual backbone and a classification head.

cation network to obtain classification probabilities.

$$o_P = \mathbf{W}_f(\mathbf{F}_P) = \mathbf{W}_f(\mathbf{M}_{time}(\mathbf{P}_B)), \quad (1)$$

where $\mathbf{W}_f \in \mathbb{R}^{D_{time} \times D_c}$ is a fully connected layer. D_{time} is the hidden size of TimeSformer $\mathbf{M}_{time}$. D_c is the number of player classes. The player identification network is trained with a cross-entropy loss:

$$\mathcal{L}_{player} = -\frac{1}{N_b} \sum_{n=1}^{N_b} \sum_{i=1}^{D_c} y_{n,i} \cdot \log(p_{n,i}), \quad (2)$$

where N_b is the batch size. $y_{n,i}$ and $p_{n,i}$ are the true label and predicted probability for sample n and class i .

3.2. Bidirectional Semantic Interaction Module

The BSIM is designed to strengthen the association between player identity-related visual features and overall video content through information exchange, ensuring that player identities are more effectively integrated into the visual context. BSIM involves a multi-head self-attention module, a multi-head cross-attention module, and an MLP layer. The BSIM aims to associate video and player features and we define this operation for the multi-head self-attention mechanism with the input vector $\mathbf{I}$:

$$\text{MSA}(\mathbf{I}) = \delta \left(\frac{\mathbf{I}\mathbf{W}_Q \cdot (\mathbf{I}\mathbf{W}_K)^\top}{\sqrt{D}} \right) \cdot \mathbf{I}\mathbf{W}_V, \quad (3)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are learnable query, key, and value parameter matrices, respectively. D is the dimension of the input vector $\mathbf{I}$. $\delta(\cdot)$ is the softmax function. The multi-head cross-attention mechanism is the variant of the multi-head self-attention mechanism, where the query comes from one input vector $\mathbf{I}_1$ while the key and value come from another input vector $\mathbf{I}_2$:

$$\text{MCA}(\mathbf{I}_1, \mathbf{I}_2) = \delta \left(\frac{\mathbf{I}_1\mathbf{W}_Q \cdot (\mathbf{I}_2\mathbf{W}_K)^\top}{\sqrt{D}} \right) \cdot \mathbf{I}_2\mathbf{W}_V. \quad (4)$$

The BSIM is shown in Fig. 4. The visual features of the k player sequences are concatenated into a player feature $\mathbf{F}_k \in \mathbb{R}^{k \times D_{time}}$, which is then interacted with the video feature $\mathbf{V}_v$ through the bidirectional semantic interaction module. Specifically, the video feature $\mathbf{V}_v$ and player feature $\mathbf{F}_k$ each pass through a multi-head self-attention layer for self-enhancement:

$$\begin{cases} \mathbf{V}'_v = \text{MSA}_1(\ell_1(\mathbf{V}_v)) \\ \mathbf{F}'_k = \text{MSA}_2(\ell_2(\mathbf{F}_k)) \end{cases}, \quad (5)$$

where $\ell(\cdot)$ denotes the Layer Normalization layer.

Then $\mathbf{V}'_v$ and $\mathbf{F}'_k$ are associated through an information exchange sub-module:

$$\begin{cases} \mathbf{V}''_v = \mathbf{W}_{u1} \text{MCA}_{ev} \left(\ell_3 \left(\mathbf{W}_{d1} \mathbf{V}'_v \right), \ell_4 \left(\mathbf{W}_{d2} \mathbf{F}'_k \right) \right) \\ \mathbf{F}''_k = \mathbf{W}_{u2} \text{MCA}_{ve} \left(\ell_4 \left(\mathbf{W}_{d2} \mathbf{F}'_k \right), \ell_3 \left(\mathbf{W}_{d1} \mathbf{V}'_v \right) \right) \end{cases}, \quad (6)$$

where $\mathbf{W}_d \in \mathbb{R}^{D_{time} \times D_{down}}$ and $\mathbf{W}_u \in \mathbb{R}^{D_{down} \times D_{time}}$ denote the linear down- and up-projection matrices. $\text{MCA}_{ev}(\cdot)$ and $\text{MCA}_{ve}(\cdot)$ denote entity-to-video multi-head cross-attention and video-to-entity multi-head cross-attention, respectively.

Finally, the outputs of BSIM are as follows:

$$\begin{cases} \mathbf{V}_{bsi} = \ell_5 \left(\varrho_1 \left(\text{MLP}_1 \left(\phi_1 \left(\mathbf{V}''_v \right) \right) \right) + \mathbf{V}''_v \right) \\ \mathbf{F}_{bsi} = \ell_6 \left(\varrho_2 \left(\text{MLP}_2 \left(\phi_2 \left(\mathbf{F}''_k \right) \right) \right) + \mathbf{F}''_k \right) \end{cases}, \quad (7)$$

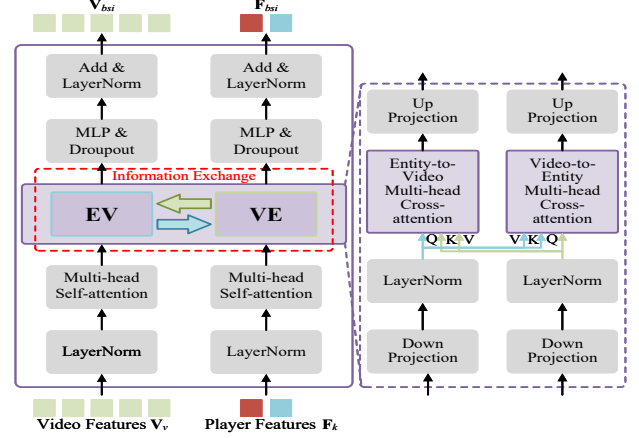


Figure 4. The bidirectional semantic interaction module.

where $\varrho(\cdot)$ denotes the Dropout layer. $\phi(\cdot)$ denotes the GELU activation function [14]. And $\text{MLP}(\cdot)$ denotes the multilayer perceptron layer.

3.3. Visual Context Learning Module

The visual context learning module (VCLM) aims to learn the video context information $\mathbf{V}_c \in \mathbb{R}^{N_q \times D_{time}}$. As shown in Fig. 2, we randomly initialize N_q learnable query vectors $\Theta \in \mathbb{R}^{N_q \times D_{time}}$. VCLM extracts a kind of “global” and “overview” semantic information from the original video features through learnable query vectors to help the model capture the overall dynamic background of the video content. Learnable vectors not only provide contextual information, but also act as a bridge between the visual and textual spaces. These vectors can be automatically adjusted during training to better fit the semantic mapping between vision and text. We use self- and cross-attention to aggregate video context information:

$$\mathbf{V}_{self} = \delta \left(\frac{\Theta \mathbf{W}^{Q_1} \cdot (\Theta \mathbf{W}^{K_1})^\top}{\sqrt{D_{time}}} \right) \cdot \Theta \mathbf{W}^{V_1}, \quad (8)$$

$$\mathbf{V}_{pv} = \tau_p + \mathbf{V}_v, \quad (9)$$

$$\mathbf{V}_{cross} = \delta \left(\frac{\mathbf{V}_{self} \mathbf{W}^{Q_2} \cdot (\mathbf{V}_{pv} \mathbf{W}^{K_2})^\top}{\sqrt{D_{time}}} \right) \cdot \mathbf{V}_{pv} \mathbf{W}^{V_2}, \quad (10)$$

$$\mathbf{V}_c = \text{FFN}(\text{MLP}(\mathbf{V}_{cross})), \quad (11)$$

where $\mathbf{V}_v \in \mathbb{R}^{N_v \times D_{time}}$ is the video feature extracted by TimeSformer [3]. N_v is the number of frames. τ is the position embedding and helps the model distinguish the sequence of frames and capture temporal dependencies. $\mathbf{W}^{Q_1} \in \mathbb{R}^{D_{time} \times D_{time}}$, $\mathbf{W}^{K_1} \in \mathbb{R}^{D_{time} \times D_{time}}$, $\mathbf{W}^{V_1} \in \mathbb{R}^{D_{time} \times D_{time}}$, $\mathbf{W}^{Q_2} \in \mathbb{R}^{D_{time} \times D_{time}}$, $\mathbf{W}^{K_2} \in \mathbb{R}^{D_{time} \times D_{time}}$ and $\mathbf{W}^{V_2} \in \mathbb{R}^{D_{time} \times D_{time}}$ are learnable matrices. $\text{FFN}(\cdot)$ denotes the feed-forward network.

3.4. Large Language Model Based Decoder

We use a pre-trained LLM as the decoder. First, the pre-trained LLM is employed to extract the corresponding textual feature $\mathbf{E}_n \in \mathbb{R}^{1 \times D_{llm}}$ for each player’s name in the player sequence. Then, we concatenate k name features into $\mathbf{E}_k \in \mathbb{R}^{k \times D_{llm}}$. To enhance generalization and adaptability, a multimodal prompt is provided to the LLM, which includes video context information $\mathbf{V}_c$, video feature $\mathbf{V}_{bsi}$, player feature $\mathbf{F}_{bsi}$, and player name feature $\mathbf{E}_k$. This prompt enables the model to generate captions $\hat{\mathbf{C}}$ that accurately incorporate player identities.

$$\hat{\mathbf{C}} = \Psi_{llm}([\mathbf{W}_c \mathbf{V}_c, \mathbf{W}_v \mathbf{V}_{bsi}, \mathbf{W}_f \mathbf{F}_{bsi}, \mathbf{W}_n \mathbf{E}_k]), \quad (12)$$

where $\mathbf{W}_c \in \mathbb{R}^{D_{time} \times D_{llm}}$, $\mathbf{W}_v \in \mathbb{R}^{D_{time} \times D_{llm}}$, $\mathbf{W}_f \in \mathbb{R}^{D_{time} \times D_{llm}}$ and $\mathbf{W}_n \in \mathbb{R}^{D_{llm} \times D_{llm}}$ are linear projection layers that map each feature into the embedding space of the LLM. D_{llm} is the hidden size of the LLM. $\Psi_{llm}(\cdot)$ denotes the LLM-based decoder, and $[\cdot]$ denotes the concatenate function in Python.

3.5. Training and Inference

During the training stage, the video_id and game_id are utilized to query the player-centric video clip set, retrieving the corresponding key player sequences and player names. The player identification network extracts player visual features from player sequences, while player names are encoded into textual features using an LLM. Following the conventional training pattern for caption generation, we utilize cross-entropy loss to optimize our LLM-IAVC:

$$\mathcal{L}_\Phi = - \sum_{t=1}^{N_g} \log(\mathbf{P}(\beta_t^* | \beta_{0:t-1}^*, \mathbf{F}_{prompt}; \Phi)), \quad (13)$$

$$\mathbf{F}_{prompt} = [\mathbf{W}_c \mathbf{V}_c, \mathbf{W}_v \mathbf{V}_{bsi}, \mathbf{W}_f \mathbf{F}_{bsi}, \mathbf{W}_n \mathbf{E}_k], \quad (14)$$

where $\{\beta_0^*, \beta_1^*, \dots, \beta_{N_g}^*\}$ is the set of ground-truth tokenized tokens. N_g denotes the number of tokens. And Φ denotes the optimized parameters.

At the inference stage, there are some differences compared to the training stage. We utilize the pre-trained multi-object tracking model SportsMOT [8] to extract multiple player sequences. These sequences are then processed by the player identification network to recognize player identities. The identified sequences are ranked based on confidence scores, and the top- k sequences are selected. The visual features of these top- k players are concatenated to form the player feature representation $\mathbf{F}_k$. Additionally, the names of the top- k players are encoded into textual features using a pre-trained LLM, resulting in a concatenated feature representation $\mathbf{E}_k$. Finally, as in Eq. (12), this concatenated feature serves as input prompt to the LLM, generating the corresponding description.

Dataset	Domain	Bbox	Video	Sentence	One	Multi.	Total Dur(h)
Anonymous							
MSVD [6]	Open	No	1.9k	70k		✓	5.3
ActivityNet-Caption [4]	Open	No	20k	100k		✓	840
MSR-VTT [49]	Open	No	10k	200k		✓	41.2
TACoS [35]	Cooking	No	127	11.8		✓	-
YouCook [9]	Cooking	No	88	2.7k		✓	2.3
YouCook2 [55]	Cooking	No	2k	15.4		✓	176
SVCDV [30]	Volleyball	No	4.8k	44k	✓		-
Soccer captioning [11]	Soccer	No	22k	22k	✓		-
SoccerNet-Caption [24]	Soccer	No	0.9k	36k	✓		715.9
MatchTime [33]	Soccer	No	0.9k	36k	✓		715.9
SoccerReplay-1988 [34]	Soccer	No	150k	150k	✓		3,323.0
FSN [51]	Basketball	No	2k	6.5k	✓		-
BH-Commentary [54]	Basketball	No	4.3k	4.3k	✓		10.1
NSVA [45]	Basketball	Yes	32k	44.6k	✓		84.8
Identity-aware							
MPII-MD [37]	Movie	No	68.3k	68.3k	✓		73
Movie101 [52]	Movie	No	30k	30k	✓		92
Goal [29]	Soccer	No	8.9k	22k	✓		25.5
VC-NBA-2022 [47]	Basketball	No	3.9k	3.9k	✓		-
NBA-Identity	Basketball	Yes	9.7k	9.7k	✓		8.9

Table 1. Comparison of existing video captioning datasets. “Bbox” denotes whether the dataset includes entity bounding boxes. “One” means each video is paired with only one caption, while “Multi.” means multiple different captions.

4. Experiments

LLM-IAVC is compared with existing video captioning models on NBA-Identity and VC-NBA-2022 [47] datasets. We further conduct ablation studies to assess the contribution of each component in LLM-IAVC.

4.1. Datasets and Metrics

We collect 40 games data from the NBA 2022-2023 season to construct an identity-aware basketball video captioning dataset **NBA-Identity**. It includes 9,726 videos, 321 players with bounding boxes, and 9 major kinds of events (e.g., “block”, “foul”, “defensive rebound”, “offensive rebound”, “turnover”, “two-point (2-pt) shot”, “three-point (3-pt) shot”, “layup” and “assist”). Notably, “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”. Our NBA-Identity covers 26 fine-grained event types. And it is divided into training and testing sets based on game instances. 8,667 clips from 35 games for training, 1,059 clips from 5 games for testing. As shown in Tab. 1, open-domain datasets (including cooking) typically provide multiple captions per video clip, whereas sports and movie datasets usually have only one. This may be due to: 1) the higher difficulty and cost of obtaining domain-specific annotations, and 2) the uniqueness of terminology in specialized fields like sports. To facilitate identification, we also provide key player bounding boxes.

VC-NBA-2022 is an identity-aware basketball video captioning dataset, which includes 9 types of fine-grained shooting events, 286 players, and over 3.9k videos. The training set contains 3,162 videos and the testing set contains 786 videos. We employ the typical captioning met-

Model	Learning-rate	Hidden size	Batch size
GPT-2 [31]	5e-5	768	64
Qwen2.5-0.5B [40]	7e-5	896	32
Qwen2.5-1.5B [40]	5e-5	1,536	32
Qwen2.5-3B [40]	1e-5	2,048	8
Llama3.2-1B [23]	5e-5	2,048	32
Llama3.2-3B [23]	7e-6	3,072	8

Table 2. Training parameters for different variants of LLM-IAVC.

rics, including BLEU [27], Rouge-L [20], METEOR [2] and CIDEr [43] to evaluate the performance of LLM-IAVC.

Compared to VC-NBA-2022, Our NBA-Identity extends VC-NBA-2022 with the following key differences: 1) More video clips; 2) More event types; 3) More realistic video clips: Video clips of NBA-Identity are in MP4 format with variable lengths based on event duration, while clips in VC-NBA-2022 are uniformly sampled to 72 frames. 4) All captions in NBA-Identity are directly obtained from live broadcast platforms without any modifications, making them authentic, while the textual annotations in VC-NBA-2022 have been manually revised into structured subject-verb-object formats. 5) Captions include shot distance, making generation more challenging and realistic. **More details about the proposed dataset and metrics can be found in Supplementary materials.**

4.2. Implementation Details

The hidden size D_{time} of TimeSformer is 768. Each video clip in NBA-Identity has up to 60 frames, with frame and player image sizes of 224×224 . Since descriptions of the dataset mention at most two players, the top-2 player sequences are selected for visual feature extraction. We stack 1 visual context learning module block and 1 bidirectional semantic interaction module layer for LLM-IAVC. The number of learnable query vectors is set to 32 for visual context learning module. And the D_{down} is set to 512 for bidirectional semantic interaction module. For player identification network training, the Adam [18] optimizer is used with a learning rate of 5e-5 for 50 epochs, covering 321 player identities. The number of images in player sequence is set to 20. LLM-IAVC is trained for 100 epochs on both datasets. Variants of LLM-IAVC utilize different LLMs as decoders, including GPT-2, Qwen2.5, and Llama3.2, with training parameters detailed in Tab. 2. The Adam optimizer is used across all variants, with a beam search size of 5 for decoding. All experiments run on an NVIDIA 4090 GPU. For example, with Llama3.2-3B (batch = 8), training and testing per epoch take 221.2s and 115.3s, respectively, with 0.3s per caption generation.

4.3. Results on NBA-Identity and VC-NBA-2022

Compared models. The proposed LLM-IAVC is compared with 5 advanced video captioning models, including Clip4Caption [39], SwinBERT [21], CoCap [38], Om-

niViD [44], KEANet [47] and EIKA [48]. Clip4Caption employs CLIP [32] to acquire the aligned visual-text representation for better generating the text descriptions. SwinBERT introduces Video Swin Transformer [22] to encode spatiotemporal representations from video frames. CoCap utilizes the motion vector feature, residual feature, and video feature through the action encoder. Then, the fused features are sent to the decoder to generate video captions. OmniViD is a unified generative framework that can address various video tasks, including action recognition, video captioning, video question answering, dense video captioning, and visual object tracking. KEANet and EIKA utilize the candidate player list to generate the descriptions with player identities.

Quantitative results analysis. In Tab. 3, LLM-IAVC significantly outperforms the other 6 methods on CIDEr. This is due to CIDEr’s sensitivity to unique or infrequent words like names, with accurate name prediction boosting scores. Traditional methods perform worse as they cannot directly recognize player identities from video. While KEANet and EIKA use a predefined candidate player list, they lack visual recognition, leading to less accurate player names.

LLM-IAVC includes multiple variants with different LLM decoders, such as GPT-2 [31], Qwen 2.5 [40] and Llama 3.2 [23]. Each variant takes either visual context alone or a multimodal prompt combining visual and identity features as input. Results show poor performance with visual context alone, while multimodal prompts significantly enhance generation quality. For example, adding prompts boosts the CIDEr score of LLM-IAVC (Llama 3.2-3B) by 40.9% and BLEU-4 by 4.3%, as the prompts provide richer visual context and player identity information for better video understanding. Among decoders, LLaMA outperforms Qwen and GPT-2 due to its superior architecture and pre-training process, leading to better contextual understanding and generation.

Due to resource limitations, we are unable to develop larger models. However, our approach is lightweight, allowing for easy deployment and training. These results underscore the capability of our model to generate accurate names and fine-grained actions in live text broadcast task.

Qualitative results analysis. Fig. 5 presents qualitative results on NBA-Identity and VC-NBA-2022. LLM-IAVC takes two input types: video-only and multimodal prompts. The video-only model struggles to understand player identities and actions, leading to inaccurate names, though they may resemble real names (e.g., “Justin Jackson” vs. “Jayson Tatum”). By integrating video context and player identities, the generation quality is significantly improved, enabling more accurate descriptions of player names and actions. Moreover, shot distance is crucial in practical applications. It can be observed that traditional and extra information enhanced methods fail to accurately determine

Method	NBA-Identity				VC-NBA-2022			
	CIDEr	METEOR	Rouge-L	BLEU-4	CIDEr	METEOR	Rouge-L	BLEU-4
Traditional methods								
Clip4Caption [39]	21.7	21.0	37.6	6.5	70.4	26.7	51.2	28.8
SwinBERT [21]	23.5	20.7	37.1	7.5	69.1	26.5	49.0	28.4
CoCap [38]	22.7	20.0	37.7	7.5	70.5	27.4	50.7	28.9
OmniViD [44]	39.6	21.6	37.7	8.3	71.2	27.5	50.7	29.2
Extra inforamtion enhanced methods								
KEANet [47]	50.1	22.7	38.1	8.5	138.5	28.0	54.9	32.4
EIKA [48]	55.7	22.7	38.6	9.0	140.7	29.5	56.8	36.7
LLM-IAVC without prompt								
LLM-IAVC (G)	46.8	17.5	34.0	9.2	87.4	27.7	53.8	30.5
LLM-IAVC (Q-0.5B)	54.3	17.9	33.8	11.5	88.1	27.6	54.2	30.8
LLM-IAVC (Q-1.5B)	58.8	18.8	38.5	12.3	90.0	27.5	55.2	31.0
LLM-IAVC (Q-3B)	61.8	18.2	39.0	14.7	101.2	28.0	55.5	32.2
LLM-IAVC (L-1B)	60.3	18.0	33.2	14.0	90.0	27.3	55.3	31.4
LLM-IAVC (L-3B)	66.5	18.2	34.5	14.2	103.1	28.5	55.7	32.8
LLM-IAVC with prompt								
LLM-IAVC (G)	94.1	22.0	35.4	17.0	139.0	29.0	57.8	37.2
LLM-IAVC (Q-0.5B)	87.6	21.5	36.4	16.0	139.4	29.2	55.7	36.8
LLM-IAVC (Q-1.5B)	97.1	22.1	40.0	16.7	141.5	30.1	57.4	37.8
LLM-IAVC (Q-3B)	103.5	22.9	41.2	17.7	147.8	30.7	56.8	38.3
LLM-IAVC (L-1B)	88.3	21.6	40.1	15.3	139.3	30.0	57.3	37.7
LLM-IAVC (L-3B)	105.3	23.0	42.5	18.8	150.7	30.9	57.9	38.6

Table 3. Quantitative comparison results on NBA-Identity and VC-NBA-2022 datasets. Within each unit, we denote the best performance in **RED** and the second-best performance in [BLUE](#). G, Q, and L represent GPT-2, Qwen2.5, and Llama3.2, respectively.

shooting distance. While our method doesn’t explicitly predict it in NBA-Identity, providing richer and more detailed features as prompts allows the model to approximate the real shot distance more closely (*e.g.*, “6 ft” vs. “4 ft”). Though distance estimation remains a challenge, our model is a flexible and extensible baseline that can incorporate a distance evaluation module for improved accuracy.

4.4. Ablation Study

Backbone of player identification network. We compare different backbones for the player identification network in Tab. 4, treating player identification as a classification task where each class represents a player identity. We employ Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA) metrics from the group activity recognition domain [46] to assess the performance of the player identification network. A higher MCA indicates stronger overall recognition ability across all player categories. A higher MPCA reflects a better measure of the model’s generalization ability, especially in scenarios with imbalanced player categories. The TimeSformer-based [3] player identification network achieves the highest MCA (91.40%) and MPCA (89.13%), demonstrating superior ac-

Backbone	Input size	MCA	MPCA
ResNet-18 [13]	320×180	86.46%	79.74%
ResNet-50 [13]	320×180	90.12%	85.30%
C3D [12]	320×180	50.86%	43.32%
I3D [5]	320×180	60.24%	57.20%
Vision Transformer [10]	224×224	87.12%	80.06%
TimeSformer [3]	224×224	91.40%	89.13%

Table 4. Impact of different backbones on player classification.

Model	VCLM	PIN	BSIM	CIDEr	BLEU-4
①				66.5	14.2
②	✓			66.8	15.0
③		✓		95.1	16.4
④	✓	✓		99.7	17.7
⑤		✓	✓	103.4	18.1
⑥	✓	✓	✓	105.3	18.8

Table 5. Ablation on NBA-Identity.

curacy and balanced player recognition.

Contribution of each component. As shown in Tab. 5, ablation experiments assess the contributions of the visual context learning module (VCLM), player identification network (PIN), and bidirectional semantic interaction module (BSIM). Without any components, model ① performs

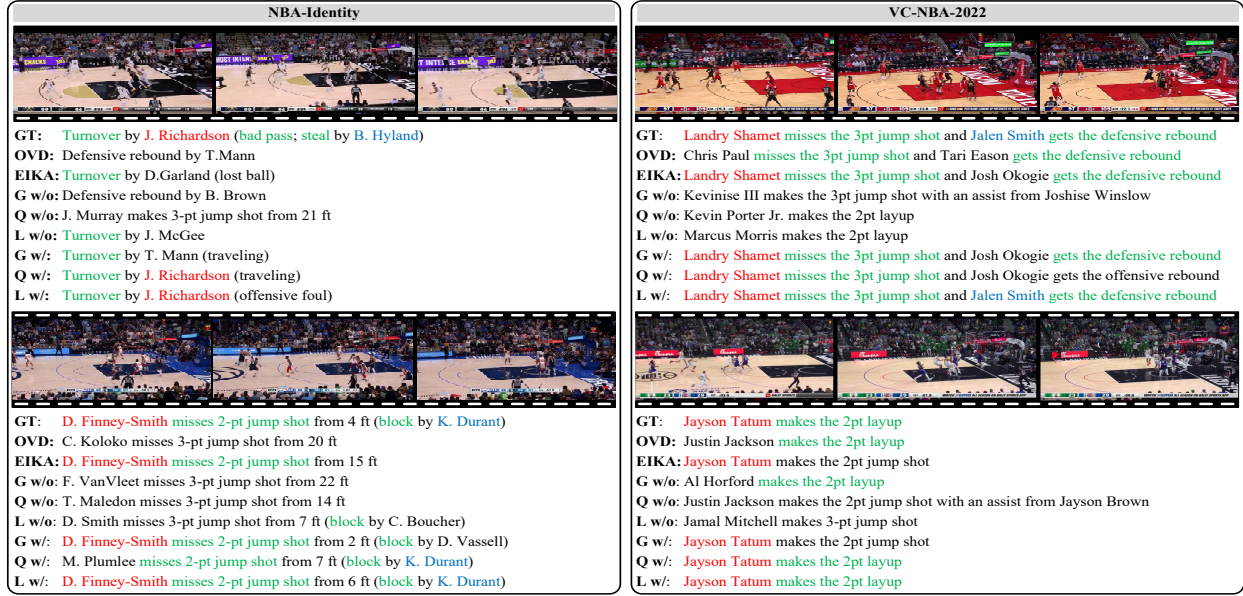


Figure 5. Qualitative results on NBA-Identity and VC-NBA-2022 datasets. “Model w/o” refers to the model using only video as input, without a prompt. “Model w/” refers to the model with a multimodal prompt as input. G, Q, and L denote GPT-2, Qwen2.5-3B, and Llama3.2-3B, respectively. Different entity names are marked **red** and **blue**. And the specific visual scenes are marked **green**.

Model	Video	Player	CIDEr	BLEU-4
①	✓		101.7	18.0
②		✓	102.3	18.3
③	✓	✓	105.3	18.8

Table 6. Ablation on the output of BSIM.

poorly due to the challenge of identifying players in complex sports scenes. Adding VCLM (model ②) slightly improves performance by helping the model focus on key video regions instead of scanning all frames. While contextual information enhances feature connections, the limited improvement reflects the difficulty of recognizing player identities from visual content.

Model ③ extends Model ① by incorporating the PIN, significantly boosting performance (CIDEr: 66.5 → 95.1). Based on this, model ④ further improves performance by leveraging key video content. Video content provides details on player actions and positions. Integrating this with player identities helps the model accurately associate players with their actions, leading to more precise descriptions. As a result, Model ⑤ significantly outperforms Model ③. Model ⑥ integrates all components, achieving the best performance. These results confirm the effectiveness of each module, highlighting the crucial role of visual-based player identification in identity-aware sports video captioning. We further explore the impact of the BSIM module outputting only a single enhanced feature (either video or player) on model performance. In Tab. 6, BSIM performs worse when outputting only one enhanced feature compared to using

both. The enhanced player features contribute the most to overall performance.

5. Conclusion, Limitation and Future Work

We propose a player-centric multimodal prompt generation network for identity-aware basketball video captioning. An identity-related information extraction module is designed to help the LLM recognize player identities and generate descriptions with player names. To support this work, we construct a large identity-aware basketball video captioning dataset containing 9,726 videos, 321 players with bounding boxes, and 9 major event types. Extensive experiments validate the effectiveness of our method.

As an initial exploration of player recognition for identity-aware sports video captioning, our work still has some limitations. Firstly, the model heavily relies on the performance of the player identification network. Future work should fully utilize the player-related knowledge, such as player numbers and jersey colors, to improve identity recognition accuracy. Secondly, the model struggles to infer shot distances, and the distance evaluation module should be added to address this issue. Thirdly, the dataset needs to include more event types to make it more comprehensive. Despite these limitations, we believe that recognizing player identities from a visual perspective is a key breakthrough for identity-aware sports video captioning and offers valuable insights. Furthermore, the proposed dataset provides valuable resources for future research.

Acknowledgments

This project was supported in part by the Beijing Natural Science Foundation under grant L233008; in part by the Natural Science Foundation of China under Grant 62236010.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [2] Satantjeev 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. 6
- [3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 2, 3, 4, 7
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 2, 5
- [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. 7
- [6] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2, 5
- [7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2
- [8] 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. 5
- [9] Pradipto Das, Chenliang Xu, Richard F Doell, et al. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013. 2, 5
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 7
- [11] Ahmad Hammoudeh, Bastien Vanderplaetse, and Stéphane Dupont. Soccer captioning: dataset, transformer-based model, and triple-level evaluation. *Procedia Computer Science*, 210:104–111, 2022. 5
- [12] Kensho Hara, Hirokatsu Kataoka, et al. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 2, 7
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 7
- [14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4
- [15] Harsh Jhamtani, Varun Gangal, Eduard Hovy, et al. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1661–1671, 2018. 2
- [16] Chen Jiang, Kaiming Huang, Sifeng He, et al. Learning segment similarity and alignment in large-scale content based video retrieval. In *MM '21: ACM Multimedia Conference*, pages 1618–1626. ACM, 2021. 1
- [17] Byeong Jo Kim and Yong Suk Choi. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1056–1065, 2020. 2
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [19] Junnan Li, Yongkang Wong, Qi Zhao, et al. Video storytelling: Textual summaries for events. *IEEE Trans. Multim.*, 22(2):554–565, 2020. 1
- [20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6
- [21] Kevin Lin, Linjie Li, Chung-Ching Lin, et al. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 1, 2, 6, 7
- [22] Ze Liu, Jia Ning, Yue Cao, et al. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6
- [23] Meta-llama. Llama 3.2 quantized models, 2024. 2, 6
- [24] Hassan Mkhallati and Anthony others Cioppa. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5085, 2023. 1, 2, 5
- [25] Liqiang Nie, Leigang Qu, Dai Meng, et al. Search-oriented micro-video captioning. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3234–3243. ACM, 2022. 1
- [26] Övgü Özdemir and Erdem Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571, 2024. 2
- [27] 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. 6
- [28] Jae Sung Park, Trevor Darrell, and Anna Rohrbach. Identity-aware multi-sentence video description. In *Computer*

- Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 360–378. Springer, 2020. 1, 2
- [29] Ji Qi, Jifan Yu, Teng Tu, et al. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5391–5395, 2023. 1, 2, 5
- [30] Mengshi Qi, Yunhong Wang, Annan Li, et al. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633, 2019. 2, 5
- [31] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 6
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 6
- [33] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation. *arXiv preprint arXiv:2406.18530*, 2024. 1, 2, 5
- [34] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards universal soccer video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8384–8394, 2025. 1, 2, 5
- [35] Michaela Regneri, Marcus Rohrbach, et al. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 5
- [36] Shuhuai Ren, Linli Yao, Shicheng Li, et al. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2
- [37] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, et al. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 5
- [38] Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, and Libo Zhang. Accurate and fast compressed video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15558–15567, 2023. 1, 2, 6, 7
- [39] Mingkan Tang, Zhanyu Wang, Zhenhua Liu, et al. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 1, 2, 6, 7
- [40] Qwen Team. Qwen2.5: A party of foundation models, 2024. 2, 6
- [41] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [43] 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. 6
- [44] Junke Wang, Dongdong Chen, Chong Luo, et al. Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18209–18220, 2024. 1, 2, 6, 7
- [45] Dekun Wu, He Zhao, Xingce Bao, et al. Sports video analysis on large-scale data. In *European Conference on Computer Vision (ECCV)*, pages 19–36. Springer, 2022. 1, 5
- [46] Lifang Wu, Meng Tian, Ye Xiang, et al. Learning label semantics for weakly supervised group activity recognition. *IEEE Transactions on Multimedia*, 2024. 7
- [47] 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, 2, 5, 6, 7
- [48] Zeyu Xi, Ge Shi, Haoying Sun, et al. Eika: Explicit & implicit knowledge-augmented network for entity-aware sports video captioning. *Expert Systems with Applications*, page 126906, 2025. 1, 2, 6, 7
- [49] Jun Xu, Tao Mei, Ting Yao, et al. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 5
- [50] Hanhua Ye, Guorong Li, Yuankai Qi, et al. Hierarchical modular network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17939–17948, 2022. 1, 2
- [51] Huanyu Yu, Shuo Cheng, Bingbing Ni, et al. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6006–6015, 2018. 5
- [52] Zihao Yue, Qi Zhang, Anwen Hu, et al. Movie101: A new movie understanding benchmark. *arXiv preprint arXiv:2305.12140*, 2023. 5
- [53] Luca Zanella, Willi Menapace, Massimiliano Mancini, et al. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024. 2
- [54] Benhui Zhang, Junyu Gao, and Yuan Yuan. A descriptive basketball highlight dataset for automatic commentary generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10316–10325, 2024. 2, 5
- [55] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 5