Neurocomputing 619 (2025) 129177

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A simple yet effective knowledge guided method for entity-aware video captioning on a basketball benchmark

Zeyu Xi [a], Ge Shi [a], Xuefen Li [a], Junchi Yan [b], Zun Li [a], Lifang Wu [a,*], Zilin Liu [a], Liang Wang [c]

[a] Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China
[b] School of Artificial Intelligence, and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, China
[c] Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, 100045, China

ARTICLE INFO

Communicated by C. Cusano

Keywords:
Entity-aware video captioning
Knowledge guided method
Basketball benchmark

ABSTRACT

Despite the recent emergence of video captioning models, how to generate the text description with specific entity names and fine-grained actions is far from being solved, which however has great applications such as basketball live text broadcast. In this paper, a new basketball benchmark for entity-aware video captioning is proposed. Specifically, we construct a multimodal basketball game knowledge graph (KG_NBA_2022) storing basketball game records as well as detailed information on teams and players. Then, a multimodal basketball game video captioning (VC_NBA_2022) dataset that contains 9 types of fine-grained shooting events and 286 players' knowledge (i.e., images and names) is constructed based on KG_NBA_2022 in an automatic approach. We also develop a simple yet effective knowledge guided entity-aware video captioning network (KEANet) based on a candidate player list in an encoder–decoder form for basketball live text broadcast. The temporal contextual information in video is encoded by introducing the Bi-directional Gated Recurrent Unit (Bi-GRU) module. And the entity-aware module is designed to model relationships among players and emphasize key players. Extensive experiments on multiple sports benchmarks demonstrate that KEANet effectively leverages additional knowledge and outperforms advanced video captioning models.

1. Introduction

Video captioning (VC) is a crucial computer vision task that requires the model to output corresponding text descriptions based on a given video. By associating visual information and text elements, this task has been blossoming and gaining increasing attention owing to its many promising applications, including automatic video title generation [1], visually-impaired assistance [2,3], video storytelling [4] and online video search [5,6].

Despite the recent rapid development in video captioning, existing methods [7–16] often struggle in real-world scenarios as they fail to generate text descriptions with specific entity names and fine-grained actions. As the basketball live text broadcast example in Fig. 1, the basketball video involves multi-person actions and complex scenes, which pose significant challenges to model's performance and generalization. Note that conventional models can only generate the simple sentence to describe the video from a macroscopic perspective (e.g., a man fails to make a shot and another man gets the rebound). In contrast, if the model has game-related knowledge, such as players who appear in the game and fine-grained actions in basketball, it can generate a knowledge-grounded text description (e.g., Brandon Ingram misses the 2pt jump shot and Justise Winslow gets the defensive rebound). In addition, existing common used benchmarks, including MSVD [17], YouCook [18], MSR-VTT [19], and ActivityNet Captions [20], simplify the task by using indefinite pronouns like "a man", "a woman" or "a group of men" instead of specific entity names. And the actions in the annotated captions are coarse-grained. These benchmarks cannot provide relevant knowledge beyond videos and fine-grained action annotations to develop models to generate text descriptions with specific entity names and fine-grained actions.

An increasing number of researchers have discovered that conventional methods and benchmarks for video captioning fail to meet the requirements of practical applications, and have made numerous attempts to address this issue. Mkhallati et al. [21] publicly release SoccerNet-Caption, the first dataset for dense video captioning in soccer broadcast videos. Although this work focuses on entity-aware video captioning, it cannot provide any additional knowledge for the

* Corresponding author.
E-mail addresses: Xzy12345@emails.bjut.edu.cn (Z. Xi), shige@bjut.edu.cn (G. Shi), lixuefen@emails.bjut.edu.cn (X. Li), yanjunchi@sjtu.edu.cn (J. Yan), zunli@bjut.edu.cn (Z. Li), lfwu@bjut.edu.cn (L. Wu), liu_zilin0113@emails.bjut.edu.cn (Z. Liu), wangliang@nlpr.ia.ac.cn (L. Wang).

https://doi.org/10.1016/j.neucom.2024.129177
Received 4 June 2024; Received in revised form 30 November 2024; Accepted 7 December 2024
Available online 16 December 2024

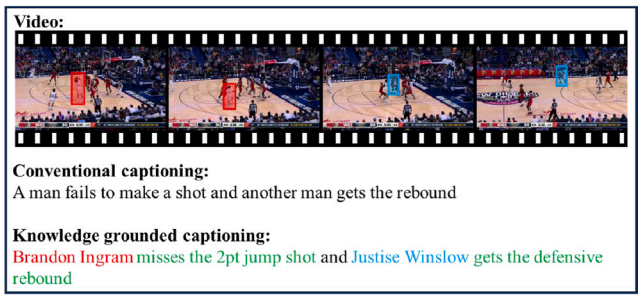


Fig. 1. Comparison of conventional captioning with knowledge-grounded captioning. Different specific entity names are marked red and blue, respectively. And fine-grained actions are marked green.

model to generate text descriptions with specific entity names. Ayyubi et al. [22] first train an entity perception detector to detect entities in video frames. They then utilize a large language model to retrieve relevant knowledge and enrich the visual content, thereby generating entity-aware text descriptions for news video summary. However, this approach is limited by the performance of entity perception detector and large language model. Furthermore, Qi et al. [23] construct a unimodal knowledge graph and a benchmark (Goal) for soccer commentary. The provided knowledge is not comprehensive and lacks visual aspects. It can be seen that these methods are tailored to different domains. Introducing relevant background knowledge is crucial for generating text descriptions with specific entity names in a particular field.

In this work, we focus on the task of knowledge guided entity-aware basketball video captioning (KEBVC), which requires the model to comprehend the given video and generate text descriptions that include specific entity names and fine-grained actions based on the additional game-related knowledge. We propose a new multimodal knowledge graph supported video captioning benchmark for basketball live text broadcast. Specifically, we collect 25 NBA full-games multimodal data from professional basketball platforms in 2022–2023 season, including events, player and team information, as well as videos. To facilitate further processing, we pre-process and structure the collected data, presenting it as a multimodal knowledge graph, as depicted in Fig. 2. The knowledge graph not only helps in organizing and presenting the information systematically but also leverages its relationships and nodes to automatically extract relevant data, forming a high-quality dataset. Subsequently, a multimodal basketball game video captioning dataset named VC_NBA_2022 is constructed based on nodes in KG_NBA_2022 and relationships among the selected nodes with an automatic approach. VC_NBA_2022 dataset comprises 9 types of basketball shooting events and 286 players' knowledge (i.e., images and names), with data samples illustrated in Fig. 3.

We utilize the relationships within the knowledge graph to obtain game-related knowledge, specifically the players who participate in the games corresponding to the videos. To effectively utilize game-related knowledge in the VC_NBA_2022 dataset, we propose a simple yet effective knowledge guided entity-aware network (KEANet) based on a candidate player list in an encoder–decoder form for basketball live text broadcast. KEANet is comprised of 3 separate unimodal encoders for videos, players' images, and players' names, as well as a pre-trained language model that serves as the text decoder for generating descriptions of the given videos. Moreover, a Bi-GRU [24] module is introduced to encode temporal contextual information, while an entity-aware module is designed to model the associations among candidate players and emphasize key players.

The main contributions of this paper are as follows:

- We provide an in-depth analysis to discover characteristics of basketball domain and construct a multimodal basketball game knowledge graph. This knowledge graph stores records of basketball games, as well as information about teams and players in an organized manner. Based on nodes and relationships in knowledge graph, a multimodal basketball game video captioning benchmark is constructed in an automatic approach.
- We develop a simple yet effective knowledge guided entity-aware video captioning network based on a candidate player list in an encoder–decoder form, which attentively incorporates key information from the additional knowledge to generate text descriptions with specific entity names.
- To validate the generalization of the proposed model, we conduct experiments on the proposed basketball dataset and the football domain dataset Goal [23]. The proposed model outperforms existing advanced models, achieving leading performance.

2. Related works

2.1. Computer vision with knowledge graph

A knowledge graph (KG) is essentially a large-scale semantic network that consists of entities and concepts as nodes, with various semantic relationships among them as edges [25]. The application of knowledge graphs have greatly promoted the rapid development of computer vision. Zhuo et al. [26] utilize the mined category-attribute relationships in a knowledge graph and the similarity between seen categories and unseen ones for more reliable knowledge transfer. One of the main challenges in the video–text retrieval task is identifying fine-grained semantic associations between video and text. To address this issue, under the guidance of additional knowledge, Fang et al. [27] utilize associations between concepts to extend new concepts and enrich the representation of videos. This approach enables a more accurate matching of video and text, leading to improved retrieval results. Gu et al. [28] propose a text-knowledge graph augmented transformer, which integrates additional knowledge and leverages multimodal information to mitigate the challenges posed by long-tail words.

In this work, we extract data from knowledge graph by utilizing relationships and nodes to automatically construct the basketball video captioning dataset. In addition, we utilize the relationships in the knowledge graph to provide each video with game-related knowledge, such as a list of players who appear in each game, thereby helping the model generate text descriptions with entity names.

2.2. Video captioning

Video captioning is a crucial task in video understanding, where models generate text descriptions for given videos. Some works [9,16,29–36] utilize a visual encoder to extract video representations from a set of video frames, and a language decoder transfer these representations to the corresponding descriptions. Recently, CLIP (Contrastive Language-Image Pre-training) [37] has demonstrated its superior performance on various visual-linguistic tasks relying on large-scale contrastive pre-training with image-text pairs. Therefore, Clip4Caption [14] employs CLIP to acquire the aligned visual-text representation, leading to significant improvements in video captioning performance. Ren et al. [38] first incorporate object counting into remote sensing image captioning, making the generated text descriptions more specific and accurate. Inspired by meta-learning [39], Yang et al. [40] extract meta-features from natural and remote sensing image classification tasks, transferring this prior knowledge to remote sensing visual captioning. This approach effectively addresses the scarcity of annotated data in remote sensing visual captioning. Benefiting from the flexibility of the transformer architecture [41], SwinBERT [8] introduces the Video Swin Transformer [42] as the video encoder to encode spatial–temporal representations from video frames. None of the previously mentioned works can address the challenge of generating action entity

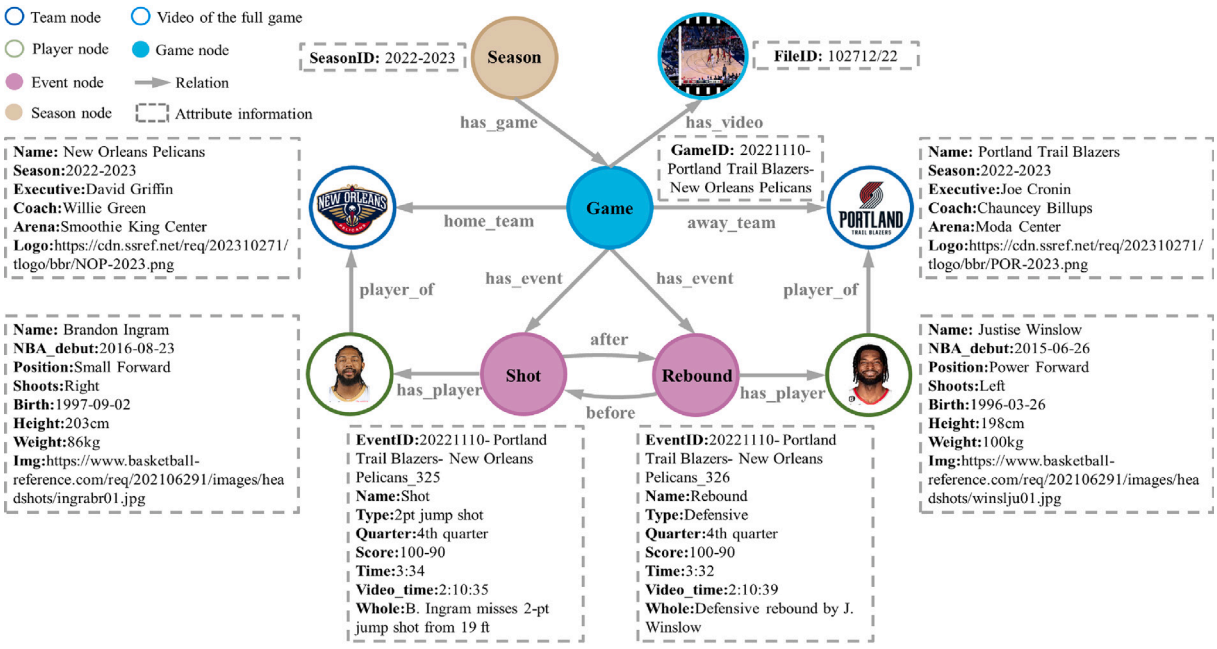


Fig. 2. An example of a multimodal basketball game knowledge graph.

Fig. 3. Data sample from the proposed dataset. Each video is annotated by fileid, action type, caption, player images and player names. Each of players involved in caption as well as their teammates serve as candidate players.

names effectively. In stark contrast, we propose to incorporate additional game-related knowledge based on an encoder–decoder model structure to generate text descriptions with specific entity names and fine-grained actions for basketball live text broadcasts. This demonstrates the practicality of our entity-aware video captioning in real applications.

2.3. Video captioning benchmarks

Existing widely used benchmarks that support the conventional video captioning task include MSVD [17], MSR-VTT [19], YouTube [43] and VATEX [44]. These benchmarks are open-domain databases that enhance the generalization of models. However, they suffer from common limitations: the text descriptions are too concise, ignoring the names of specific entities and fine-grained action types. Some works have attempted to enrich these benchmarks and develop models to generate text descriptions with more fine-grained information. Fang et al. [45] construct a benchmark annotated with captions and commonsense descriptions. This commonsense benchmark develops models to generate captions, as well as 3 types of commonsense descriptions: intention, effect, and attribute. Yu et al. [46] propose a fine-grained sports narrative benchmark that focuses on the detailed actions of subjects, but this benchmark ignores the names of specific subjects. To assist visually impaired individuals in enjoying movies, Yue et al. [47] construct a large-scale Chinese movie benchmark, which requires models to generate role-aware narration paragraphs when there are no actors speaking. Byeong et al. [48] propose a benchmark for automatically generating commentary on baseball games. The descriptions in this benchmark focus on player categories rather than specific player names. To achieve meaningful news summarization, Ayyubi et al. [22] propose the task of summarizing news video directly to entity-aware captions. They also release a large-scale dataset to support research on this task. Mkhallati et al. [21] release a benchmark for soccer broadcasts commentaries. Although the text descriptions in the benchmarks contain specific player names, they do not provide any experimental support for generating captions with specific names. On the contrary, each specific name is replaced by a specific token (e.g., [TEAM], [COACH], [REFEREE], and [PLAYER]).

Qi et al. [23] utilize the unimodal knowledge graph for real-time soccer commentary, incorporating all pre-match and player-team information without a selection mechanism. This indiscriminate inclusion could introduce noise, affecting caption accuracy. Contrastingly, our approach first employs a multimodal knowledge graph, integrating both text and visual information, such as player names and images. This enriches the context for video captioning, improving the match with video content. Second, we propose an entity-aware module, modeling

Table 1
Statistics of the events in KG_NBA_2022.

Events	Foul	Rebound	Violation	Timeout	Free throw	Enter the game	Turnover	Jump ball	Shot
num.	986	**2649**	68	271	1103	1189	774	49	**4400**

"num." is the abbreviation of "number". Numbers in bold indicate the highest two values.

associations among candidate players and incorporating key players' knowledge. This effectively reduces noise and enhances caption precision. These advancements distinguish our work from Qi et al.'s [23] and other existing models, demonstrating the potential of multimodal knowledge and entity-aware in video captioning tasks.

2.4. Spatial–temporal feature modeling

In the field of video understanding, effectively modeling spatial–temporal features is crucial for capturing the complex relationships between the spatial and temporal dimensions of video data. Video data is inherently dynamic, with objects moving across frames and evolving over time, making it essential to model both the spatial information (e.g., object locations, shapes) and temporal dependencies (e.g., motion) to understand the content of a video.

Recent developments have introduced more sophisticated appro aches to spatial–temporal feature modeling. Yang et al. [49,50] propose the spatial–temporal attention to help model simultaneously consider temporal context and spatial details, thereby generating detailed and accurate descriptions. Temporal dynamics modeling is a key challenge in video generation. Therefore, Fei et al. [51] investigate how to enhance diffusion models' awareness of video dynamics to achieve high-quality text-to-video generation. Fei et al. [52] introduce a fine-grained structural spatial–temporal alignment framework that enhances video-language models by utilizing novel graph transformers for spatial–temporal video feature propagation and introducing a spatial–temporal Gaussian differential graph transformer to capture object changes across. Yang et al. [53] improve person re-identification performance by extracting and integrating features from a spatial–temporal perspective. Park et al. [54] model spatial–temporal information by utilizing the spatial–temporal forward and backward SSM mechanism [55], which captures both non-sequential spatial and sequential temporal relationships in video. This approach allows the model to efficiently process videos while effectively capturing long-range dependencies across both space and time. Zhang et al. [56] propose a multi-object tracker based on spatial–temporal topological constraints to address challenges like irregular motion patterns, similar appearances, and frequent occlusions. Specifically, it introduces a feature adaptive association module to establish spatial–temporal associations between motion and appearance, enabling complementary integration of appearance and motion features for more accurate tracking.

It is obvious that spatial–temporal feature modeling is crucial for video understanding tasks. In this paper, we employ the convolutional neural network ResNet-18 [57] to extract local spatial feature from each frame and utilize the Bi-GRU [24] to model the temporal context across frames, thereby achieving more accurate video content understanding.

3. Proposed benchmark

In this section, we first construct a multimodal knowledge graph to store the records of games and the information of teams and players. Then, we utilize relationships and nodes in the graph to automatically extract relevant data and construct a multimodal basketball dataset. This automated construction method reduces the heavy costs of laborious manual annotations.

3.1. Data collection and pre-processing

Our knowledge graph is based on basketball game videos with corresponding event descriptions. To begin constructing the multimodal basketball game knowledge graph, we collect full-game play-by-play data from 50 games, including event descriptions, the time on the scoreboard corresponding to the events, score records, player information, and team information from a professional basketball data platform.[1] The corresponding videos are collected from a basketball broadcast platform.[2] After filtering out videos with low resolution and chaotic content, only 25 games are retained. To structure the collected data, following steps are needed to process it: (1) Categorize basketball events. (2) Parse and structure the descriptions for each type of event. (3) Match event descriptions with their corresponding video timestamps.

By analyzing the play-by-play data, game events can be divided into 9 categories, including "Foul", "Rebound", "Violation", "Timeout", "Free throw", "Enter the game", "Turnover", "Jump ball", and "Shot". Each text description has its owns keyword, such as in the description of the "Foul" event, "Personal foul by G. Temple (drawn by A. Drummond)", where the keyword is "foul". Therefore, the type of this event is "Foul". For this semi-structured data, the sentence is parsed into multi-tuples using the character index of the keyword and specific words that appear in the sentence, such as "drawn by". To incorporate video information into the multi-tuples, the video timestamp is intended to be associated with the text description. OCR (Optical Character Recognition) is employed to identify the time on the scoreboard in each frame and record the timestamp of the current frame. Specifically, we employ the open-source OCR toolkit PaddleOCR[3] and Tesseract-OCR [58] to recognize the time simultaneously. By using multiple OCR toolkits, we can improve the overall accuracy of recognition. This is because one toolkit may succeed where the other fails. In addition to event information, player and team information are also stored in multi-tuples. Through these efforts, the collected data is structured into a multimodal basketball game knowledge graph containing 11 489 events and 42 870 relationships. Fig. 2 shows an example of the multimodal basketball game knowledge graph (KG_NBA_2022).

3.2. Multimodal basketball video captioning dataset

To construct a dataset containing videos, text descriptions, and additional game-related knowledge (i.e., player images and names), the pertinent data is extracted from the knowledge graph by utilizing rich relations. In basketball, shot and rebound events are the most prevalent. Meanwhile, we count the number of different types of events in the knowledge graph. Table 1 shows that there are 4400 shot events and 2649 rebound events, which are the most common types of events. Based on the aforementioned common knowledge and statistical data, videos and event descriptions about shot and rebound are extracted from KG_NBA_2022 to construct the dataset. To be more realistic, the text descriptions need to be further modified.

When a shot fails to score in basketball, it is often followed by a rebound event. In line with this pattern, we have combined the shot and rebound into a single event. Shots in basketball can be categorized as two-point (2pt) shots, three-point (3pt) shots, and layups. Rebounds are further divided into defensive (def.) rebounds and offensive (off.)

[1] https://www.basketball-reference.com
[2] https://fishkernba.com
[3] https://github.com/PaddlePaddle/PaddleOCR

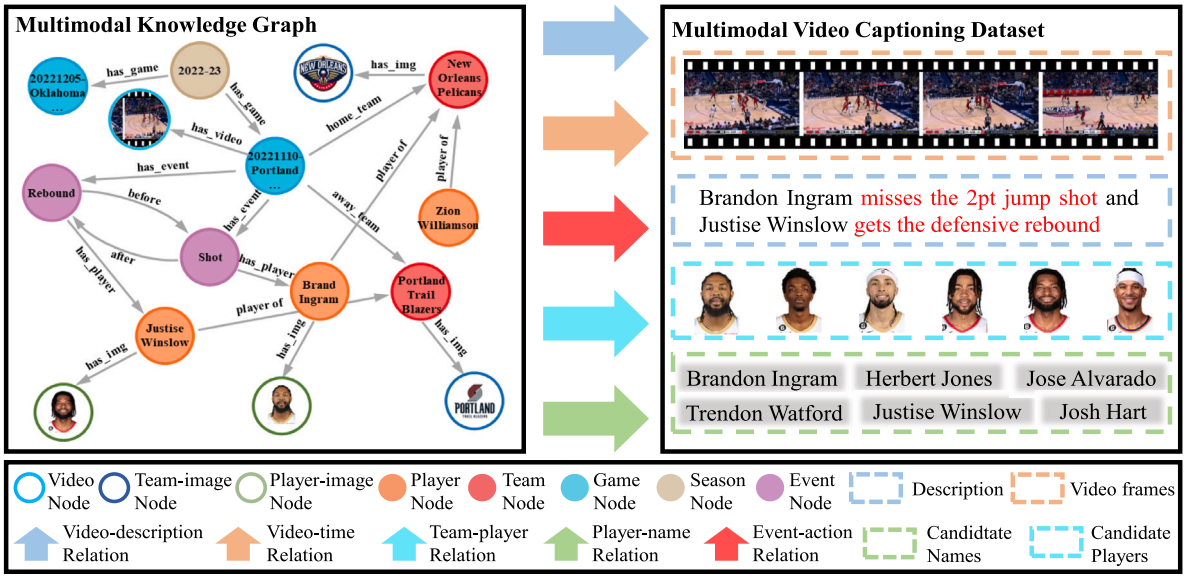


Fig. 4. Illustration of extracting relevant data using relationship extraction from the knowledge graph and constructing the dataset.

Table 2
Statistics of the labels in VC_NBA_2022.

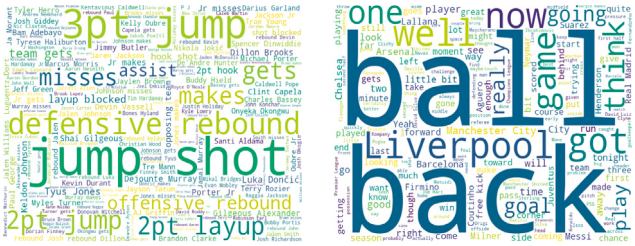
Labels	2p -succ.	2p -fail.-off	2p -fail.-def.	2p -layup -succ.	2p -layup -fail.-off.	2p -layup -fail.-def.	3p -succ.	3p -fail.-off.	3p -fail.-def.
Train num.	469	146	397	442	133	251	470	202	652
Test num.	95	41	95	108	32	67	125	61	162

"succ." and "fail." are abbreviations of "success" and "failure", respectively.

Table 3
Comparisons of different video caption datasets.

Dataset	Sentences per second	Verbs per sentence	Verb ratio
MSR-VTT	0.067	1.37	14.8%
YouCook	0.056	1.33	12.5%
ActivityNet Captions	0.028	1.41	10.4%
VC_NBA_2022	**0.182**	**1.73**	**14.8%**

Numbers in bold indicate the highest value.



(a) VC_NBA_2022 dataset (b) Goal dataset

Fig. 5. Word cloud of VC_NBA_2022 and Goal datasets. The bigger the font, the more percentage it occupies.

rebounds. The classification scheme of our dataset, based on the NBA dataset [59], is utilized for group activity recognition. However, unlike the NBA dataset, where all videos are 6 s long and divided into 72 frames, we consider that different events may have varying durations. Consequently, in our dataset, videos of different lengths are uniformly divided into 72 frames. Additionally, we have incorporated text descriptions and player information into our dataset.

As shown in Fig. 4, the event-action relation in the graph is utilized to extract 9 types of events, and the video-time relation in the graph is utilized to obtain the video timestamp of each event. Specifically, we first roughly estimate the start and end timestamps of the event based on the existing timestamp. Each clip undergoes a manual review process to ensure accurate start and end timestamps. Subsequently, the video-description relation is utilized to obtain the text description, which is then matched with the corresponding video clip. We extract player-related information (player images and player names) from KG_NBA_2022 through the team-player and player-name relations. Since each event involves certain players, only the involved individuals and their teammates need to be considered as candidate players for video annotation.

Taking the sentence "B. Ingram misses 2-pt jump shot from 19 ft and defensive rebound by J. Winslow" as an example, we will explain our text modification process. Due to the difficulty in generating distance in video captioning task, the "from 19 ft" in the sentence is removed. Based on the attribute information of the player in KG_NBA_2022, the abbreviated name ("J. Winslow") in the sentence is replaced with the full name ("Justise Winslow"). This change might increase the complexity of name generation but is more aligned with real-world scenarios. To improve text fluency, alterations are made based on the character index. Specifically, we identify the index of keywords like "defensive" and apply a rule ("defensive/offensive rebound by SOMEONE" $\Longrightarrow$ "SOMEONE gets the defensive/offensive rebound") to revise the sentence. Finally, there are a total 3977 videos,[4] each of which belongs to one of 9 types of events. Each video has one text description and several candidate players information (images and names). We randomly select 3162 clips for training and 786 clips for testing. We also provide word-cloud-based statistics in Fig. 5(a) to reveal the relative amount of different words. It shows that the top-4

[4] The scale of the current dataset is not large enough, which is emergent to be enriched in our subsequent research.

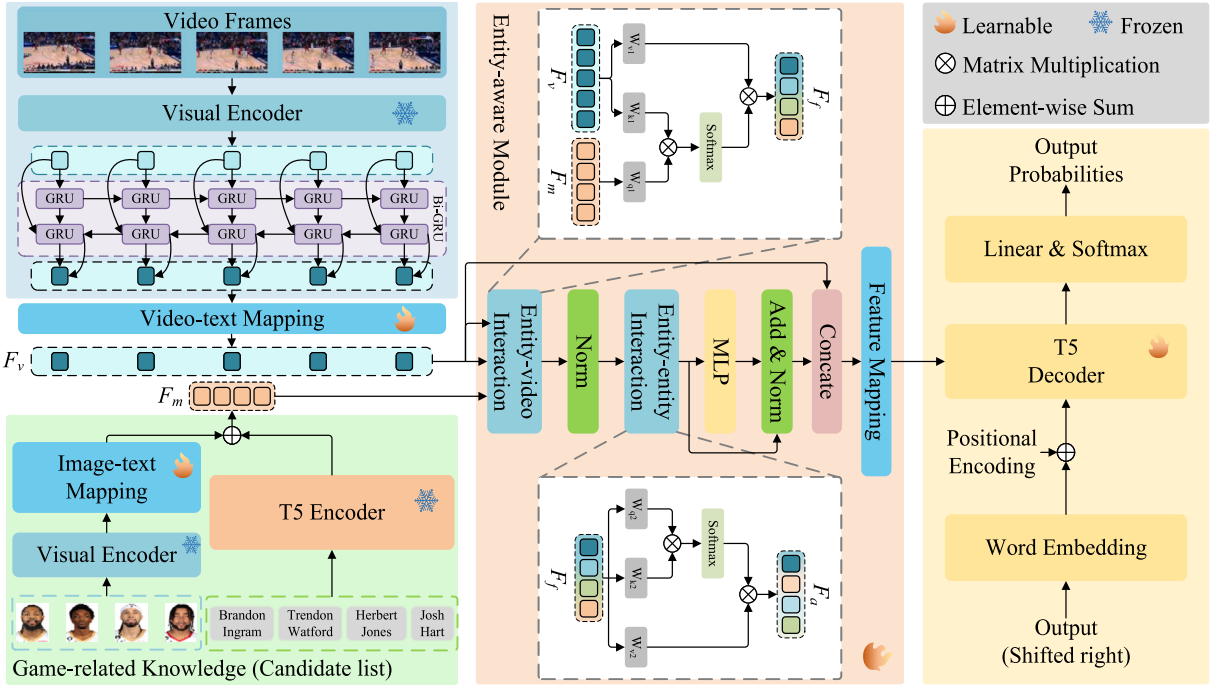


Fig. 6. The architecture of knowledge guided entity-aware video captioning network (KEANet) based on a candidate player list in an encoder–decoder form.

subjects in VC_NBA_2022 are "jump", "shot", "3pt", and "defensive", followed by "rebound", "2pt", "layup", and "makes". Table 2 shows the sample distributions across different labels of events.

In Table 3, the comparison of our dataset with MSR-VTT, YouCook and ActivityNet Captions further demonstrates the fine-grained details of our captioning annotations. VC_NBA_2022 has the most sentences per second of 0.182, while the other datasets are all below 0.1. This indicates that our dataset contains more detailed information in its descriptions. Moreover, VC_NBA_2022 has 1.73 verbs in a sentence on average, higher than 1.41 for ActivityNet Captioning and 1.37 for MSR-VTT. Similarly, the verb ratio of VC_NBA_2022, computed by dividing the total number of verbs by the total number of words in the sentence, is also significantly higher than that of the other three datasets. This highlights that our dataset primarily focuses on the fine-grained actions of the subjects, aligning with our original intent.

In real-world scenarios, the number of NBA players is limited. Our dataset contains 25 games from 23 teams, 3948 video clips, and a total of 286 players. Our task focuses on entity-aware video captioning in the sports domain, where text descriptions include not only player names but also player actions (e.g., three-point shots, two-point shots) and interactions between players (e.g., assists, blocks). Additionally, venues and lighting conditions vary across different games. All of the above factors increase the diversity of the data.

4. Knowledge guided entity-aware basketball video captioning

4.1. Problem formulation

To generate precise and concise text descriptions for basketball live text broadcast, we introduce the concept of KEBVC task. In this task, the model is required to comprehend the content of the provided video and generate text descriptions that include specific entity names and fine-grained actions based on the additional knowledge. This task can be formulated as: given the basketball video V_b, the objective is to select video-related player knowledge $K_p = s\left(V_b\right)$ and generate the text description $D_{v,k}$.

$$\text{KEBVC}: D_{v,k} = m\left(V_b, K_p\right) = m\left(V_b, s\left(V_b\right)\right), \tag{1}$$

where $s(\cdot)$ denotes the model's abilities on aligning video and knowledge (player information). $m(\cdot)$ transfers video and knowledge to text description.

4.2. Proposed model

For the KEBVC task, we propose a simple yet effective knowledge guided entity-aware video captioning network (KEANet) based on a candidate player list in an encoder–decoder form. The overall structure of KEANet framework is shown in Fig. 6. Given the raw video frames which are of size $T \times 3 \times H \times W$, consisting of T frames and each has $3 \times H \times W$ pixels, we feed them into a CNN-based visual encoder in KEANet and extract the video feature $F_T \in \mathbb{R}^{T \times D_r \times h \times w}$. D_r is the hidden size of visual feature. The global representation of F_T is then fed into Bi-GRU module to further encode temporal contextual information and obtain the feature $F_v \in \mathbb{R}^{T \times D}$.

$$F_v = W_1\left(\text{GRU}\left(\text{AvgPool}\left(F_T\right)\right)\right), \tag{2}$$

where $W_1 \in \mathbb{R}^{D_r \times D}$ is the linear mapping layer, which maps the video feature to text space. D is the hidden size of the decoder module. $\text{GRU}(\cdot)$ denotes the Bi-GRU module and $\text{AvgPool}(\cdot)$ denotes the average pooling layer.

Each video has corresponding N candidate players' information as the additional knowledge to assist in generating a text description with specific entity names. Each player image is of size $3 \times H_p \times W_p$. For candidate players' images, we employ a CNN-based visual encoder to extract features $F_N \in \mathbb{R}^{N \times D_r \times h_p \times w_p}$. The global features $F_p \in \mathbb{R}^{N \times D}$ of N images are obtained by (3).

$$F_p = W_2\left(\text{AvgPool}\left(F_N\right)\right), \tag{3}$$

where $W_2 \in \mathbb{R}^{D_r \times D}$ is a linear mapping layer, which maps the image feature to text space.

The text encoder of large language model T5 [60] is employed to transform candidate names into a sequence of embeddings $T_n \in \mathbb{R}^{N \times D}$. After that, the image features F_p and name features T_n are added by corresponding positions to obtain multimodal player features $F_m \in \mathbb{R}^{N \times D}$.

Table 4
Combined inference performance on VC_NBA_2022.

Model	CIDEr	METEOR	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Eps
V2C	13.7	18.7	45.9	50.1	39.5	27.0	14.9	0.0
Clip4Caption	70.4	26.7	51.2	49.1	42.5	35.4	28.8	20.9
Clip4Caption[a]	117.2	27.5	52.0	50.5	43.7	37.0	29.4	28.3
SwinBERT	69.1	26.5	49.0	47.8	41.4	34.5	28.4	20.7
SwinBERT[a]	120.1	27.9	52.1	51.3	43.9	37.4	30.3	29.1
CoCap	70.5	27.4	50.7	49.8	42.3	35.6	28.9	21.2
CoCap[a]	122.3	28.3	52.9	52.2	44.6	38.0	30.3	29.0
OmniViD	71.2	27.5	50.7	49.7	42.4	36.5	29.2	22.6
OmniViD[a]	125.2	28.6	53.3	52.5	45.2	38.7	30.5	30.5
KEANet	**138.5**	**28.0**	**54.9**	**53.1**	**46.4**	**38.8**	**32.4**	**31.0**

Numbers in bold indicate the best performance.
[a] Denotes the model with the candidate player list.

Basketball events involve interactions between multiple players. For example, in "Myles Turner makes the 2pt layup with an assist from Tyrese Haliburton", "Myles Turner" scores a two-point layup with the assist of "Tyrese Haliburton". Therefore, we design an entity-aware module to emphasize key players from a candidate player list. The architecture of entity-aware module is shown in Fig. 6. Entity-aware module first utilizes the entity-video interaction sub-module to fuse player features and video features to obtain F_f. This connects the players to the video content, putting the players in a specific scene.

$$F_f = \sigma\left(\frac{F_m W_{q1} \cdot \left(F_v W_{k1}\right)^{\mathrm{T}}}{\sqrt{D}}\right) \cdot F_v W_{v1}, \tag{4}$$

where $W_{q1} \in \mathbb{R}^{D\times D}$, $W_{k1} \in \mathbb{R}^{D\times D}$ and $W_{v1} \in \mathbb{R}^{D\times D}$ are learnable matrices. $\sigma(\cdot)$ denotes the softmax function.

F_f is then sent to the entity-entity interaction sub-module to model the relationship among candidate players.

$$F_f = \sigma\left(\frac{\ell\left(F_f\right) \cdot W_{q2} \cdot \left(F_f W_{k2}\right)^{\mathrm{T}}}{\sqrt{D}}\right) \cdot \ell\left(F_f\right) \cdot W_{v2}, \tag{5}$$

where $W_{q2} \in \mathbb{R}^{D\times D}$, $W_{k2} \in \mathbb{R}^{D\times D}$ and $W_{v2} \in \mathbb{R}^{D\times D}$ are learnable matrices. $\ell(\cdot)$ denotes the Layer Normalization.

The player features based on attention are concatenated with the video features and subsequently fed into the T5 decoder $\Psi_{\mathrm{T5}}(\cdot)$ to generate the text description C.

$$C = \Psi_{\mathrm{T5}}\left(W_3\left(\mathrm{Concat}\left[\ell\left(\mathrm{MLP}\left(F_a\right) + F_a\right), F_v\right]\right)\right), \tag{6}$$

where $W_3 \in \mathbb{R}^{D\times D}$ is a linear mapping layer, which maps the concatenated feature to the vector space of T5 model. $\mathrm{Concat}[,]$ denotes the concatenation function in Python, and $\mathrm{MLP}(\cdot)$ denotes the MLP layer.

The Bi-GRU module, 3 linear layers, entity-aware module and text decoder are trained to maximize the log-likelihood over the training set given by (7).

$$\mathcal{L}_\Theta = \sum_{t=1}^{N_C} log\mathrm{P}_{\mathrm{r}}\left(y_t | y_{t-1}, \left[F_m, F_v\right]; \Theta\right), \tag{7}$$

where y_t denotes the one-hot vector probability of each word at time t. N_C denotes the length of caption. And Θ denotes learnable parameters.

5. Experiments

To evaluate the performance of the proposed model, KEANet is compared with advanced video captioning models on VC_NBA_2022 and Goal [23]. We further conduct ablation experiments to verify the effectiveness of each component in KEANet.

5.1. Implementation details

The proposed KEANet utilizes ResNet-18 [57] pre-trained on the ImageNet dataset [61] as the visual encoder, whose hidden size D_r is 512. KEANet is trained on the proposed VC_NBA_2022 and Goal datasets with 100 training epochs. In these datasets, T frames are sampled by using segment-based method [62]. A high resolution of video frame is helpful for the model to understand the players' action in the video. Each frame size is 1280×720. T is set to 18. The size of player's image is 180×120 and the number of candidate players in each video is not fixed. For the large language model T5, the hidden size is 768. It is worth noting that the parameters of the visual encoder and text encoder are frozen during training. During the training stage, KEANet is optimized by ADAM [63] with the learning rate of 3e-5 and weight decay of 1e-4. Beam search with a beam size of 2 is utilized for inference. The proposed KEANet is implemented with Python 3.7 and PyTorch 1.12, and is performed on a server with an Nvidia 3090ti GPU.

5.2. Evaluation metrics

Existing video captioning benchmarks mainly adopt ngram-based metrics, including CIDEr [64], METEOR [65], Rouge-L [66], and BLEU [67]. BLEU evaluates the quality of the generated text by calculating the n-gram overlap between the generated text and reference text. Rouge-L pays more attention to generating the longest common subsequence between the generated text and reference text, and evaluates sentence quality in this way. METEOR takes more factors into account, including stem matching, synonym matching, and word order. CIDEr utilizes TF-IDF [68] to assign different weights to n-grams of varying lengths, then calculates the cosine similarities of n-grams between the generated text and reference text, averaging them to obtain the final score. These metrics primarily emphasize the consistency of text rather than the accuracy of semantics. In other words, if the generated text differs from the reference in structure or vocabulary, it may be penalized by these n-gram based evaluation metrics, even if the conveyed information is completely correct. However, for entity names, they cannot be replaced with similar words. To accurately evaluate the performance of specific entity name generation, we introduce the entity-precision score (Eps), which is calculated by dividing the number of correctly predicted names by the total number of names. For Eps, we extract entity names from the ground truth and generated texts.

5.3. Results on VC_NBA_2022

KEANet is compared with five advanced video captioning models, including V2C [45], Clip4Caption [14], SwinBERT [8], CoCap [15] and OmniViD [16]. V2C is a Transformer-based model that generates relevant commonsense descriptions of the given video. Clip4Caption employs CLIP to acquire the aligned visual-text representation for better generating text descriptions. SwinBERT introduces Video Swin Transformer to encode spatial–temporal representations from video frames. CoCap fuses the motion feature, residual feature, and video feature through the encoder. Then the fused features are sent into the decoder to generate video descriptions. OmniViD is a unified generative framework that can address various video tasks, including action recognition, captioning, video question answering, dense video captioning and visual object tracking. V2C, Clip4Caption, SwinBERT and OmniViD take only videos as input. CoCap takes the motion feature, residual feature and video feature as input. And the above models are all trained and tested on our proposed VC_NBA_2022 dataset.

As shown in Table 4, KEANet outperforms other 5 models by a large margin on CIDEr. This notable performance can be attributed to the way CIDEr calculates cosine similarities of n-grams between the generated text and reference text. Highly accurate name prediction contributes to a higher cosine similarity between the n-grams of the generated text and the reference text. V2C, Clip4Caption, SwinBERT, CoCap and OmniViD exhibit lower performance in accurately generating names, resulting in much lower CIDEr scores. Although KEANet achieves the best performance across all metrics, the differences are not as pronounced in METEOR, Rouge-L, and BLEU. This can be attributed to the concise nature of the text descriptions in the dataset, primarily

Fig. 7. Qualitative results on VC_NBA_2022 dataset. (V2C: Video2Commonsense; C4C: Clip4Caption; SwB: SwinBERT; CCP: CoCap; OVD: OmniViD; KEA: our proposed model; GT: the ground truth). Different specific entity names are marked in red and blue, respectively. And fine-grained actions are marked green. Since V2C does not have its own tokenizer and vocabulary list, it cannot decode names. So, names are replaced with the special token <UNK>.

consisting of names and fine-grained actions. Notably, KEANet outperforms CoCap 9.8% on Eps and outperforms OmniViD 8.4% on Eps. The higher Eps indicates better name prediction. These results underscore the capability of our model to generate accurate names and fine-grained actions in live text broadcast task. Moreover, The model's performance significantly improves when augmented with additional knowledge. For instance, Clip4Caption[a] achieves a 48.8% increase in CIDEr and a 7.4% improvement in Eps. Similarly, OmniViD[a] improves OmniViD's CIDEr by 54.0% and Eps by 7.9%. The above experimental results show the effectiveness of additional knowledge (candidate player list) for entity-aware video captioning.

Fig. 7 shows the qualitative results on VC_NBA_2022 dataset, including generation results of V2C, Clip4Caption, SwinBERT, CoCap, OmniViD and our KEANet model. V2C, Clip4Caption, SwinBERT, CoCap and OmniViD can correctly generate some actions but fail to generate correct entity names because these names never appear during training. However, with the help of the additional game-related knowledge, our KEANet could well relate entities in video clips with the fine-grained actions. These cases demonstrate that our newly proposed KEANet model more consistently with the requirement of practical application.

Further, we study the relation between the metric CIDEr and Eps. To this end, we plot the performance of Clip4Caption (C4C), SwinBERT (SwB), CoCap (CCP), OmniViD (OVD) and KEANet (KEA) with different Eps in Fig. 8. The plot shows that this relationship is approximately linear: more accurate entity name generation imply better model performance.

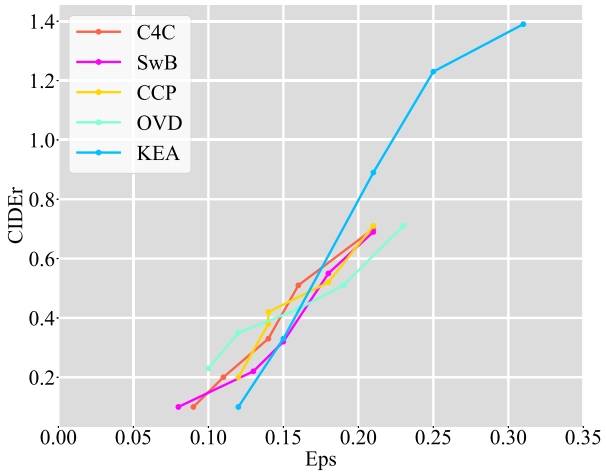


Fig. 8. The relationship between metric CIDEr and Eps.

5.4. Results on goal

Goal is a benchmark which contains over 8.9k soccer video clips, 22k sentences and 42k knowledge triples. These sentences are converted from the commentator's audios. This type of comment sentences have a certain degree of colloquialism and are relatively long. Therefore, this dataset poses certain challenges. On this basis, we filter out sentences and videos that do not contain entity names. We modify the format of the dataset to be the same as VC_NBA_2022. In addition, we save the names of teams and players in the dataset. The revised Goal dataset is shown in Fig. 9.

KEANet is compared with five advanced video captioning models, including V2C, Clip4Caption, SwinBERT, CoCap and OmniViD. V2C, Clip4Caption, SwinBERT and OmniViD take only videos as input. CoCap takes the motion feature, residual feature and video feature as input. And they are all trained and tested on the revised Goal dataset. As shown in Table 5, we compare the performance of V2C, Clip4Caption, SwinBERT, and KEANet on several metrics including CIDEr, METEOR, Rouge-L, and BLEU-1. KEANet is supported by additional knowledge that it can generate text descriptions with entity names. Therefore, the KEANet outperforms other models in all metrics. When additional knowledge is added to Clip4Caption, SwinBERT, CoCap and OmniViD, their performance is improved. This indicates that the introduction of additional knowledge (candidate player list) is beneficial for entity aware video captioning task.

Fig. 10 shows the qualitative results on Goal dataset, including generation results of V2C, Clip4Caption, SwinBERT, CoCap, OmniViD and our KEANet. Compared to the other five models, KEANet can generate more correct entity names and partial actions. The above results indicate that, even in challenging tasks such as sports commentary, with the support of additional knowledge, the model can generate text with entity names.

5.5. Ablation study

To verify the contributions of the additional knowledge and other modules in KEANet, we also perform an ablation study by progressively adding these as input. The results of ablation study are shown in Table 6. Model ① consists of Resnet-18 and T5 decoder, with its input being solely video. Model ② adds the Bi-GRU module on top of model ①. The Bi-GRU models the temporal contextual relationships between video frames, enabling the model to better understand the dynamic information in videos and predict more accurate action categories. Therefore, model ② has shown improvements in all metrics relative to model ①, except for the Eps score. From the comparison

Table 5
Combined inference performance on Goal.

Method	CIDEr	METEOR	Rouge-L	BLEU-1
V2C	0.1	2.1	3.4	3.4
Clip4Caption	2.2	5.0	5.5	5.7
Clip4Caption[a]	2.5	5.2	5.8	5.9
SwinBERT	2.2	5.1	5.3	5.7
SwinBERT[a]	2.6	5.5	6.0	5.9
CoCap	2.3	5.0	5.3	5.5
CoCap[a]	2.5	5.3	5.8	5.9
OmniViD	3.0	5.9	9.1	10.7
OmniViD[a]	**3.9**	**6.4**	**10.6**	14.4
KEANet	3.7	**6.4**	10.5	**14.9**

Numbers in bold indicate the best performance.
[a] Denotes the model with the candidate player list.

results between of model ① and model ③, it can be observed that adding players' images to model ① does not improve the performance. During the training stage, if the model has not been exposed to enough specific names, or has not learned how to infer names from video and image features, it may not be able to generate these names. The comparison results between model ① and model ④ show a significant improvement by solely adding players' names based on the model ①. Providing a list of names as input allows the model to effectively integrate this information with video features, leading to more accurate generation of names. This is likely providing the model with additional context information to help it make more accurate predictions. From the comparison results between model ④ and model ⑤, we can find that player images can emphasize the roles based on entity names and improve the names' prediction. From the comparison results of models ⑤-⑧, it can be observed that adding entity-aware and Bi-GRU modules can improve the performance. The entity-aware module can model the associations among players and focus on key players. The Bi-GRU module can model the temporal contextual information of video frame features, capturing the action information. From the above results, the additional players knowledge brings significant gains in entity awareness. The model can generate text descriptions with specific player names and fine-grained actions through additional knowledge and temporal modeling.

5.6. Impact of dynamically changing knowledge on performance of KEANet

In real-world scenarios, factors such as player trades and team recruitment can result in variations in the player roster across different games. Consequently, the dynamic nature of knowledge might impact the model's performance. To explore the extent to which this dynamic knowledge affects KEANet's performance, we conduct two experiments on the VC_NBA_2022 dataset: (1) KEANet-random: The candidate player list for each video in the training set remains unchanged, while two players in the candidate list for each video in the testing set are randomly replaced; (2) KEANet-key: The candidate player list for each video in the training set remains unchanged, while two key players in the candidate list for each video in the testing set are randomly replaced. Here, key players refer to the top-10 most frequent players in the entire dataset. By randomly replacing players in the candidate player lists, we simulate real-world scenarios where player changes (dynamic knowledge) occur.

The corresponding experimental results are shown in Fig. 11. After randomly replacing players, the KEANet's performance shows a decline. For example, the CIDEr score drops from 138.5 to 121.1 and 117.2, while the BLEU-4 score decreases from 32.4 to 31.0 and 29.2. Changes in the inputs of KEANet leads to a decline in its performance. It is well known that player changes within a team, especially within a single season, are relatively infrequent. Our dataset is collected from games during the 2022–2023 season, where knowledge variation is minimal. Furthermore, we utilize the entire roster of both teams from the video clip rather than just the players on the field, which largely ensures consistency in knowledge between the training and testing sets.

Fig. 9. Data sample from the revised Goal dataset. Each video is annotated by fileid, team_id, caption, player images and player names. Each of the players involved in caption as well as their teammates serve as candidate players. The caption in this sample includes the names of players and team.

Fig. 10. Qualitative results on Goal dataset. (V2C: Video2Commonsense; C4C: Clip4Caption; SwB: SwinBERT; CCP: CoCap; OVD: OmniViD; KEA: our proposed model; GT: the ground truth). The different specific entity names are marked in red. And the fine-grained actions are marked green. Since V2C does not have its own tokenizer and vocabulary list, it cannot decode names. So, names are replaced with the special token <UNK>.

Table 6
Ablation study on VC_NBA_2022.

Model	Ki	Kn	Ea	Bi-GRU	CIDEr	METEOR	Rouge-L	BLEU-4	Eps
①					20.1	20.6	40.9	23.0	5.5
②				✓	23.5	22.7	43.1	24.4	5.7
③	✓				18.0	20.0	38.0	23.2	10.0
④		✓			110.6	25.2	49.1	27.3	24.9
⑤	✓	✓			115.7	25.6	53.1	28.1	26.6
⑥	✓	✓	✓		119.3	26.8	53.3	29.5	27.5
⑦	✓	✓		✓	122.2	27.3	54.6	30.5	29.5
⑧	✓	✓	✓	✓	**138.5**	**28.0**	**54.9**	**32.4**	**31.0**

Ki and Kn denote the player images knowledge and player names knowledge, respectively. Ea is the entity-aware module and Bi-GRU is the bi-directional GRU module. Numbers in bold indicate the best performance.

5.7. Error analysis

Error analysis is conducted by presenting cases of KEANet on the proposed VC_NBA_2022 dataset. As shown in Fig. 12, we show several errors. (1) Player Mismatching: when the number of players in the sentence is more than one, the model may decode the name of one of the players incorrectly. (2) Action Confusion: model tends to confuse similar-looking actions, such as layups and close-range two-point jump shots. (3) Lack Distance-aware Perception: model tends to confuse three-point jump shots and long-range two-point jump shots. For example, the player appears to be shooting a three-point shot, but is actually inside the three-point line.

Given the above result, we discuss some potential ways for developing an advanced models for KEBVC task. First, beyond the current object detection, it is necessary to enhance the model's ability to understand and handle names. Second, regarding action confusion, this may be because the model is not sensitive enough to subtle differences in the video to accurately distinguish similar actions. We need to improve the feature extraction abilities of the model to more accurately recognize similar-looking actions. Third, for the lack of distance perception, this may be because the model has trouble processing spatial information, especially when judging the relative distance of players from the basket. This indicates that we need to further improve the spatial awareness of our model.

6. Conclusion, limitation and future work

In this paper, we introduce the task of knowledge guided entity-aware video captioning (KEBVC) for basketball live text broadcast. To investigate this task, we construct a multimodal basketball game knowledge graph (KG_NBA_2022) that stores records of basketball games as well as information about teams and players. Through partial relationships and nodes in the knowledge graph, a new multimodal basketball game video captioning dataset is then constructed in an automatic approach. We also propose a simple yet effective knowledge guided entity-aware video captioning network (KEANet) for generating captions with specific entity names by leveraging the game-related knowledge, i.e., a list of players who participate in the games corresponding to the videos. Without the aid of additional detection, this simple method effectively assists the model in generating players' names. Furthermore, our experiments validate the significance of incorporating additional knowledge, such as player images and entity names, in enhancing entity-aware video captioning performance.

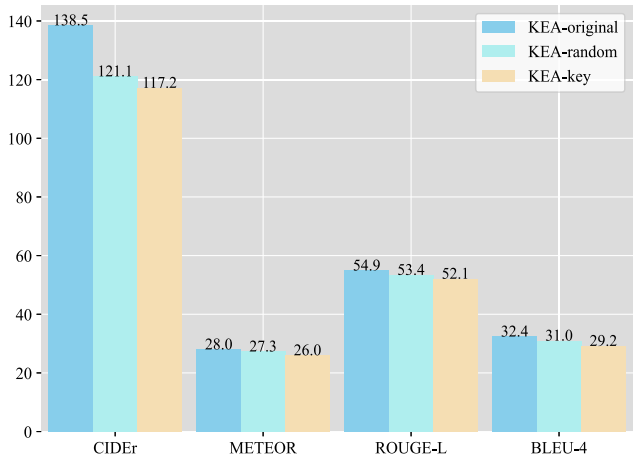


Fig. 11. Impact of dynamically changing knowledge on performance of KEANet.

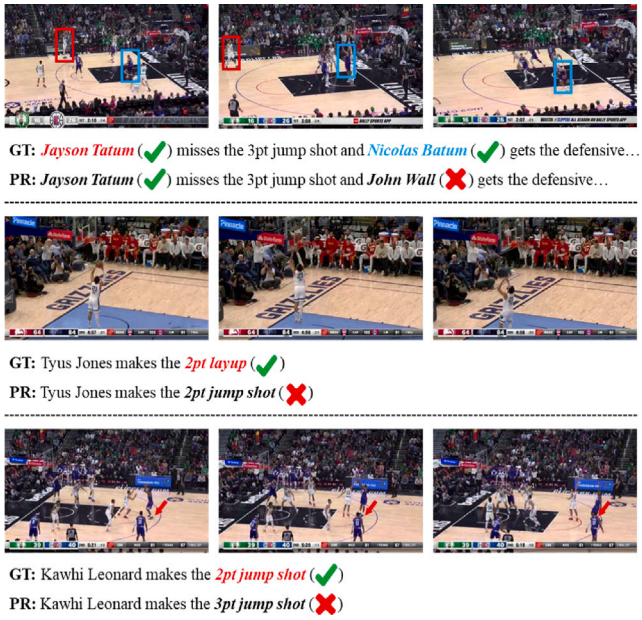


Fig. 12. Representative error cases of the generated captions, which correspond to the player mismatching, action confusion and lack distance-aware perception. GT denotes the ground truth description and PR denotes the generated description.

As an initial exploration of entity-aware video captioning for basketball live text broadcast with additional knowledge, our work still preserves several limitations that need to be improved. Firstly, due to the simplicity of the additional knowledge, although our model can generate player names, there is still significant room for accuracy improvement. Therefore, it is necessary to expand the existing knowledge graph and extract more comprehensive and informative knowledge from sports-related knowledge graphs for sports entity-aware captioning. Secondly, in future research, we aim to expand the dataset and enrich the diversity of the captions. Thirdly, the dynamic nature of knowledge changes would affect the model's performance. In future work, we will explore more effective methods to enhance the model's robustness. Fourthly, player counting has potential benefits for basketball live text broadcast, as it can provide more detailed information and enhance the contextual relevance of captions (e.g. two defenders are blocking the shooter). Future work will focus on extending and improving in this direction. Despite these limitations, we believe that incorporating the candidate player list is valuable for the task of sports entity-aware captioning. And our KG_NBA_2022 and VC_NBA_2022 can be valuable resources for numerous active researchers.

CRediT authorship contribution statement

Zeyu Xi: Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Ge Shi:** Formal analysis, Conceptualization. **Xuefen Li:** Validation. **Junchi Yan:** Writing – review & editing, Formal analysis. **Zun Li:** Conceptualization. **Lifang Wu:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition. **Zilin Liu:** Conceptualization. **Liang Wang:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledge

This work is supported in part by the Natural Science Foundation of China under Grant 62236010, 62106010, 62306022; in part by the Beijing Natural Science Foundation under grant L233008; in part by the China Postdoctoral Science Foundation, China under Grant 2022M720318; Beijing Postdoctoral Science Foundation under Grant 2022-zz-077.

Data availability

Data will be made available on request.

References

[1] Z. Zhang, Z. Ma, C. Yuan, et al., Chinese title generation for short videos: Dataset, metric and algorithm, IEEE Trans. Pattern Anal. Mach. Intell. (01) (2024) 1–16.
[2] T. Han, M. Bain, A. Nagrani, et al., Autoad III: The prequel–back to the pixels, 2024, arXiv preprint arXiv:2404.14412.
[3] Z. Yue, Y. Zhang, Z. Wang, et al., Movie101v2: Improved movie narration benchmark, 2024, arXiv preprint arXiv:2404.13370.
[4] H. Liu, J. Yang, C.-H. Chang, et al., AOG-LSTM: An adaptive attention neural network for visual storytelling, Neurocomputing 552 (2023) 126486.
[5] G. Pardi, S. Gottschling, Y. Kammerer, The influence of knowledge type and source reputation on preferences for website or video search results, J. Assoc. Inf. Sci. Technol. 75 (5) (2024) 521–537.
[6] M. Choi, H. Goel, M. Omama, et al., Neuro-symbolic video search, 2024, arXiv preprint arXiv:2403.11021.
[7] T. Mahmud, F. Liang, Y. Qing, et al., CLIP4VideoCap: Rethinking clip for video captioning with multiscale temporal fusion and commonsense knowledge, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.
[8] K. Lin, L. Li, C.-C. Lin, et al., Swinbert: End-to-end transformers with sparse attention for video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.
[9] H. Luo, L. Ji, B. Shi, et al., Univl: A unified video and language pre-training model for multimodal understanding and generation, 2020, arXiv preprint arXiv:2002.06353.
[10] T. Jin, Z. Zhao, P. Wang, et al., Interaction augmented transformer with decoupled decoding for video captioning, Neurocomputing 492 (2022) 496–507.
[11] P. Li, P. Zhang, X. Xu, Graph convolutional network meta-learning with multi-granularity POS guidance for video captioning, Neurocomputing 472 (2022) 294–305.
[12] B. Zhao, M. Gong, X. Li, Hierarchical multimodal transformer to summarize videos, Neurocomputing 468 (2022) 360–369.
[13] B. Yang, T. Zhang, Y. Zou, CLIP meets video captioning: Concept-aware representation learning does matter, in: Chinese Conference on Pattern Recognition and Computer Vision, Springer, 2022, pp. 368–381.
[14] M. Tang, Z. Wang, Z. Liu, et al., Clip4caption: Clip for video caption, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4858–4862.

[15] Y. Shen, X. Gu, K. Xu, et al., Accurate and fast compressed video captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15558–15567.

[16] J. Wang, D. Chen, C. Luo, et al., Omnivid: A generative framework for universal video understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18209–18220.

[17] D. Chen, W.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, 2011, pp. 190–200.

[18] P. Das, C. Xu, R.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, 2013, pp. 2634–2641.

[19] J. Xu, T. Mei, T. 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, 2016, pp. 5288–5296.

[20] R. Krishna, K. Hata, F. Ren, et al., Dense-captioning events in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 706–715.

[21] H. Mkhallati, A. Cioppa, S. Giancola, et al., SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 5073–5084.

[22] H.A. Ayyubi, T. Liu, A. Nagrani, et al., Video summarization: Towards entity-aware captions, 2023, arXiv preprint arXiv:2312.02188.

[23] J. Qi, J. Yu, T. 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, 2023, pp. 5391–5395.

[24] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation. arxiv 2014, 2020, arXiv preprint arXiv:1406.1078.

[25] X. Zhu, Z. Li, X. Wang, et al., Multi-modal knowledge graph construction and application: A survey, IEEE Trans. Knowl. Data Eng. (2022).

[26] J. Zhuo, Y. Zhu, S. Cui, et al., Zero-shot video classification with appropriate web and task knowledge transfer, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5761–5772.

[27] S. Fang, S. Wang, J. Zhuo, et al., Concept propagation via attentional knowledge graph reasoning for video-text retrieval, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4789–4800.

[28] X. Gu, G. Chen, Y. Wang, et al., Text with knowledge graph augmented transformer for video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 18941–18951.

[29] L. Li, J. Lei, Z. Gan, et al., VALUE: A multi-task benchmark for video-and-language understanding evaluation, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks, Virtual, 2021.

[30] X. Luo, X. Luo, D. Wang, et al., Global semantic enhancement network for video captioning, Pattern Recognit. 145 (2024) 109906.

[31] Z. Zhang, Z. Qi, C. Yuan, et al., Open-book video captioning with retrieve-copy-generate network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 9837–9846.

[32] S. Liu, Z. Ren, J. Yuan, Sibnet: Sibling convolutional encoder for video captioning, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 1425–1434.

[33] H. Luo, L. Ji, M. Zhong, et al., Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning, Neurocomputing 508 (2022) 293–304.

[34] S. Jing, H. Zhang, P. Zeng, et al., Memory-based augmentation network for video captioning, IEEE Trans. Multimedia 26 (2024) 2367–2379.

[35] P. Song, D. Guo, X. Yang, et al., Emotional video captioning with vision-based emotion interpretation network, IEEE Trans. Image Process. (2024).

[36] Y. Ma, Z. Zhu, Y. Qi, et al., Style-aware two-stage learning framework for video captioning, Knowl.-Based Syst. 301 (2024) 112258.

[37] A. Radford, J.W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[38] Z. Ni, Z. Zong, P. Ren, Incorporating object counts into remote sensing image captioning, Int. J. Digit. Earth 17 (1) (2024) 2392847.

[39] R. Vilalta, Y. Drissi, A perspective view and survey of meta-learning, Artif. Intell. Rev. 18 (2002) 77–95.

[40] Q. Yang, Z. Ni, P. Ren, Meta captioning: A meta learning based remote sensing image captioning framework, ISPRS J. Photogramm. Remote Sens. 186 (2022) 190–200.

[41] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017, pp. 5998–6008.

[42] Z. Liu, J. Ning, Y. Cao, et al., Video swin transformer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 3202–3211.

[43] M. Sun, A. Farhadi, S. Seitz, Ranking domain-specific highlights by analyzing edited videos, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 787–802.

[44] X. Wang, J. Wu, J. Chen, et al., Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4581–4591.

[45] Z. Fang, T. Gokhale, P. Banerjee, et al., Video2Commonsense: Generating commonsense descriptions to enrich video captioning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Online, November 16-20, 2020, pp. 840–860.

[46] H. Yu, S. Cheng, B. Ni, et al., Fine-grained video captioning for sports narrative, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6006–6015.

[47] Z. Yue, Q. Zhang, A. Hu, et al., Movie101: A new movie understanding benchmark, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023, pp. 4669–4684.

[48] B.J. Kim, Y.S. Choi, Automatic baseball commentary generation using deep learning, in: The 35th ACM/SIGAPP Symposium on Applied Computing, ACM, 2020, pp. 1056–1065.

[49] Y. Tu, X. Zhang, B. Liu, C. Yan, Video description with spatial-temporal attention, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1014–1022.

[50] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: Spatial-temporal attention mechanism for video captioning, IEEE Trans. Multimedia 22 (1) (2019) 229–241.

[51] H. Fei, S. Wu, W. Ji, H. Zhang, T.-S. Chua, Dysen-VDM: Empowering dynamics-aware text-to-video diffusion with LLMs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7641–7653.

[52] H. Fei, S. Wu, M. Zhang, M. Zhang, T.-S. Chua, S. Yan, Enhancing video-language representations with structural spatio-temporal alignment, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[53] X. Yang, X. Wang, L. Liu, N. Wang, X. Gao, STFE: a comprehensive video-based person re-identification network based on spatio-temporal feature enhancement, IEEE Trans. Multimed. (2024).

[54] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, Y. Qiao, Videomamba: State space model for efficient video understanding, in: European Conference on Computer Vision, Springer, 2025, pp. 237–255.

[55] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2023, arXiv preprint arXiv:2312.00752.

[56] J. Zhang, M. Wang, H. Jiang, X. Zhang, C. Yan, D. Zeng, STAT: Multi-object tracking based on spatio-temporal topological constraints, IEEE Trans. Multimed. (2023).

[57] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[58] R. Smith, An overview of the tesseract OCR engine, in: Ninth International Conference on Document Analysis and Recognition, 2, IEEE, 2007, pp. 629–633.

[59] R. Yan, L. Xie, J. Tang, et al., Social adaptive module for weakly-supervised group activity recognition, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 208–224.

[60] C. Raffel, N. Shazeer, A. Roberts, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (1) (2020) 5485–5551.

[61] O. Russakovsky, J. Deng, H. Su, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[62] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 20–36.

[63] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, 2015.

[64] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[65] S. Banerjee, A. 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, 2005, pp. 65–72.

[66] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[67] K. Papineni, S. Roukos, T. Ward, et al., Bleu: a method for automatic evaluation of machine translation, in: Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[68] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, J. Doc. 60 (5) (2004) 503–520.

Zeyu Xi received the M.S. degree in control science and engineering from Yanshan University, Qinhuangdao, China, in 2022. He is currently pursuing his Ph.D. in electronic science and technology from Beijing University of Technology, Beijing, China. His research interests include single target tracking, video captioning, multimodal large language models and multimodal knowledge graph.

Ge Shi received the Ph.D. degree in Computer Science from Beijing Institute of Technology in 2020. He is now an Associate Professor at the Faculty of Information Technology, Beijing University of Technology, China. His main research interests include information extraction, text generation, and cross-modal learning.

Xuefen Li received the B.S. degree in communication engineering from Wuhan Polytechnic University, China, in 2022. She is currently pursuing the M.S. degree in electronic information from Beijing University of Technology, China. Her current research interests include video grounding and natural language processing.

Junchi Yan was a Senior Research Staff Member at IBM Research, Beijing, China, where he started his career in April 2011. He is currently a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include machine learning and computer vision. Dr. Yan received the ACM China Doctoral Dissertation Nomination Award and the China Computer Federation Doctoral Dissertation Award for his work on graph matching. He regularly serves as a Senior PC/Area Chair for Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), Computer Vision and Pattern Recognition (CVPR), Association for the Advancement of Artificial Intelligence (AAAI), International Joint Conferences on Artificial Intelligence (IJCAI), and ACM International Conference on Multimedia (ACM MM) and an Associate Editor for the Pattern Recognition journal.

Zun Li received her Ph.D. degree from the School of Computer and Information Technology, Beijing Jiaotong University in 2021. She was a visiting scholar with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2018 to 2019. She is currently a lecturer at the Faculty of Information Technology, Beijing University of Technology, China. Her research interests include computer vision and machine learning, object detection and segmentation, and temporal action detection.

Lifang Wu received her BS, MS and Ph.D. from Beijing University of Technology in 1991, 1994, and 2003 respectively. She is currently a Professor at Beijing University of Technology. Her research interests include image/video understanding, group activity recognition, face anti-spoofing, multi-model sentiment analysis, and intelligent 3D printing. She is a CCF outstanding member. She has received 5 provincial and ministerial science and technology awards. She has also received Best Paper award of ICCV 2021 Workshop on Human-centric Trustworthy Computer Vision and PRCV 2021 Best Paper Honorable Mentions. She participated in organizing the CCCV 2017, PRCV 2019, PRCV 2022, ChinaMM 2021, ChinaMM 2020, CCIG 2022 and so on.

Zilin Liu is currently pursuing her B.S. in electronic and information engineering from Beijing University of Technology, Beijing, China. Her research interests include multimodal knowledge graph and video captioning.

Liang Wang received the B.Eng. and M.Eng. degrees from Anhui University, Hefei, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2004. From 2004 to 2010, he was a Research Assistant with Imperial College London, London, U.K., and Monash University, Clayton, VIC, Australia, a Research Fellow with the University of Melbourne, Parkville, VIC, Australia, and a Lecturer with the University of Bath, Bath, U.K. He is currently a Full Professor of hundred talents program with the National Laboratory of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He is a IAPR Fellow.