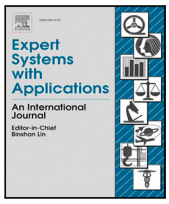
Expert Systems With Applications 274 (2025) 126906

Contents lists available at ScienceDirect

Expert Systems With Applications

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



EIKA: Explicit & Implicit Knowledge-Augmented Network for entity-aware sports video captioning

Zeyu Xi, Ge Shi, Haoying Sun, Bowen Zhang, Shuyi Li, Lifang Wu *

School of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China

ARTICLE INFO

Keywords:
Entity-aware sports video captioning
Explicit knowledge
Implicit knowledge

ABSTRACT

Sports video captioning in real application scenarios requires both entities and specific scenes. However, it is difficult to extract this fine-grained information solely from the video content. This paper introduces an Explicit & Implicit Knowledge-Augmented Network for Entity-Aware Sports Video Captioning (EIKA), which leverages both explicit game-related knowledge (i.e., the set of involved player entities) and implicit visual scene knowledge extracted from the training set. Our innovative Entity-Video Interaction Module (EVIM) and Video-Knowledge Interaction Module (VKIM) are instrumental in enhancing the extraction of entity-related and scene-specific video features, respectively. The spatiotemporal information in video is encoded by introducing the Spatial-Temporal Modeling Module (STMM). And the designed Scene-To-Entity (STE) decoder fully utilizes the two kinds of knowledge to generate informative captions with the distributed decoding approach. Extensive evaluations on the VC-NBA-2022, Goal and NSVA datasets demonstrate that our method has the leading performance compared with existing methods.

1. Introduction

Sports video captioning aims to generate a sentence that describes the main content of the sports video, which has potential applications in various real-world scenarios, such as live text broadcast (Xi et al., 2025) and commentary generation (Cook & Karakuş, 2024; Gautam et al., 2024; Mkhallati, Cioppa, Giancola, et al., 2023; Qi, Yu, Tu, et al., 2023; Zhang, Gao and Yuan, 2024). It tends to be challenging because videos usually involve complex scenes depicting the interactions among multiple player entities.

Traditional methods (Aafaq, Akhtar, Liu, et al., 2019; Lin, Li, Lin, et al., 2022; Pan, Cai, Huang, et al., 2020; Tang, Wang, Liu, et al., 2021; Xu, Yao, Zhang, et al., 2017; Yao, Torabi, & Cho, 2015; Ye, Li, Qi, et al., 2022; Zhang, Qi, Yuan, et al., 2021) aggregate a fixed set of video frames features into a video representation via an encoder, and employ a language decoder operates on top of the video representation to learn visual-textual alignment for caption generation. Although these methods achieve promising results in open domains, they tend to provide rough descriptions of the video content, overlooking key details that audiences are genuinely interested in, such as player entity names and specific scenes (Fig. 1(a)).

Recently, some methods have attempted to enhance entity awareness by incorporating explicit information, which refers to information that can be directly obtained from existing resources and databases (Qi et al., 2023; Xi et al., 2025), or visible visual information acquired through external tools (Kim & Choi, 2020; Wu, Zhao, Bao, et al., 2022). For example, Qi et al. (2023) extract game-related information from existing sports platforms, such as game news and player statistics, helping the model obtain player-related information from explicit sources to generate sports commentary with player identities. However, despite these augmentations, the quality of the captions remains suboptimal due to two primary reasons. (1) External explicit knowledge is independent of the video content, failing to provide semantic or visual cues about the current scene. This limits the model's ability to recognize different game scenarios, resulting in less accurate scene classifications in the generated descriptions. (2) The common practice of merely concatenating external explicit information with visual features before processing them through the encoder leads to inefficient utilization of information. This approach results in poor data integration and oversimplifies the complex relationships between external explicit information and visual features, thereby preventing the decoder from forming a comprehensive understanding of the combined inputs. These reasons lead to inaccurate descriptions of this type of method (Fig. 1(b)).

In fact, with the principle of analogical reasoning in Cognitive Psychology (Smith & Kosslyn, 2007), humans search their long-term

* Corresponding author.
E-mail addresses: Xzy12345@emails.bjut.edu.cn (Z. Xi), shige@bjut.edu.cn (G. Shi), sunhaoying97@emails.bjut.edu.cn (H. Sun), zhangbw2023@emails.bjut.edu.cn (B. Zhang), syli2022@bjut.edu.cn (S. Li), lfwu@bjut.edu.cn (L. Wu).

https://doi.org/10.1016/j.eswa.2025.126906
Received 29 September 2024; Received in revised form 4 February 2025; Accepted 13 February 2025
Available online 22 February 2025

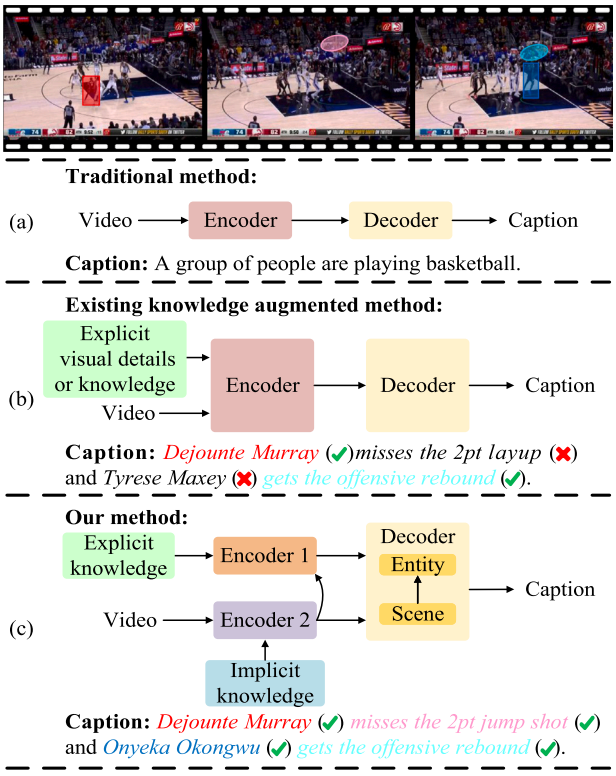


Fig. 1. Existing methods for sports video captioning VS. our proposed method. The different player entity names are marked blue and red. The specific scenes are marked pink and cyan.

memory for existing knowledge to match and understand current observations. For example, when encountering a new video, people retrieve previously formed visual scene concepts from their minds and compare the new video with these concepts to determine its category. Moreover, different videos in the identical scene category may share the certain paradigm or subtle associations of that scene. Such knowledge of visual scene concepts is implied within the video data. If the model could incorporate this implicit knowledge as the scene concept that defines and delineates different scenes, it would be possible to enhance the scene recognition capability of the model and consequently further improve the quality of the descriptions.

Motivated by the above insights, this paper focuses on mining the implicit visual scene knowledge from the video data to help the model distinguish different sports scenes and designing a framework that effectively integrates explicit and implicit knowledge to achieve a more comprehensive understanding of video content (Fig. 1(c)). Specifically, we average all video features within the same scene category in the training set to obtain visual feature centers. Each center serves as a representative description for each scene category. It is highly abstract and implicitly represents the scene knowledge, which defines the certain paradigm of one scene and also delineates the boundaries with other scenes. This implicit knowledge can guide the model in distinguishing different visual scenes. Moreover, we employ a set of player entities involved in each game as the explicit game knowledge (Xi et al., 2025) to assist model in generating descriptions with entity names. The scene and entity statistics of the training sets for the VC-NBA-2022 (Xi et al., 2025), Goal (Qi et al., 2023), and NSVA (Wu et al., 2022) datasets are shown in Fig. 2.

To effectively utilize and integrate these two kinds of knowledge, we design the Explicit & Implicit Knowledge-Augmented (EIKA) network for entity-aware sports video captioning. Specifically, our innovative video-knowledge interaction module utilizes the learnable query vectors to adaptively capture the scene-related video features with the guidance of scene knowledge, which significantly improves the depth of scene understanding. The proposed entity-video interaction module applies the attention mechanism on the game knowledge and scene-related video features to deeply analyze the interactions between players and scenes, yielding the entity-related video features. To optimally leverage both scene-related and entity-related features, we design a decoder that employs a scene-to-entity approach. It first decodes scene-related features, then decodes captions that include player names within the scene. It enhances the accuracy and coherence of generated captions by providing contextual understanding, which enables more precise alignment of scene-related entities. Additionally, the spatial–temporal modeling module is introduced to encode the spatiotemporal dynamic information by capturing both temporal and spatial relationships among video frames.

The key contributions of this paper are as follows:

- We provide an in-depth analysis to illustrate the necessity of the implicit knowledge for entity-aware sports video captioning task, and mine the scene knowledge inside the videos by a simple yet effective way.
- An Explicit & Implicit Knowledge-Augmented (EIKA) network is proposed for entity-aware sports video captioning. Our EIKA enhances the accuracy of captions by combining explicit knowledge (player names) obtained from an external knowledge graph and implicit knowledge (scene patterns) mined from the video data. Explicit knowledge provides direct entity information, while implicit knowledge helps the model gain a deeper understanding of complex scenes.
- To validate the performance of EIKA, we conduct extensive experiments on VC-NBA-2022 (Xi et al., 2025), Goal (Qi et al., 2023) and NSVA (Wu et al., 2022). Experimental results validate the effectiveness of combining the two types of knowledge and demonstrate the superior performance of EIKA.

2. Related works

2.1. Video captioning

Video captioning task requires the model to understand the spatial–temporal dynamics in video and bridge visual and textual elements to generate long sequences of output words. Recent research focuses mainly on sequence-learning generation process methods. These methods (Aafaq et al., 2019; Khan, Hussain, Ullah Khan, Ahmad Khan, & Baik, 2024; Li, Wang, Zhao, Xu, & Song, 2025; Lin et al., 2022; Pan et al., 2020; Shen et al., 2023; Tang et al., 2021; Wang et al., 2024; Wu, Song, Wang, & Zhang, 2024; Xiong et al., 2025; Ye et al., 2022; Zhang, Liu and Wu, 2024) employ a visual encoder to extract useful visual information from the given sports video, and its decoder generates the caption sequentially. Researchers have attempted to employ various visual encoders, including ResNet (He, Zhang, Ren, et al., 2016), Vision Transformer (ViT) (Dosovitskiy, Beyer, Kolesnikov, et al., 2020), SlowFast (Feichtenhofer, Fan, Malik, et al., 2019), and S3D (Miech, Alayrac, Smaira, et al., 2020), to extract different 2D/3D video features. Some efforts (Ayyubi, Liu, Nagrani, et al., 2023; Chen & Jiang, 2021; Hou, Wu, Zhang, et al., 2020; Zhang, Shi, Yuan, et al., 2020) enrich video features by capturing fine-grained static objects in videos using additional detectors. And some efforts (Ayyubi et al., 2023; Fei, Jiang, & Mao, 2021; Gu, Chen, Wang, et al., 2023; Xu, Huang, Hou, et al., 2024; Yang, Cao, & Zou, 2023; Zhang et al., 2021) also utilize retrieval augmentation, enriching video features by retrieving relevant visual or textual information from external databases to generate more accurate descriptions. Recently, large-scale pre-trained language-image models like CLIP (Radford, Kim, Hallacy, et al., 2021) have demonstrated remarkable multimodal understanding capabilities, leading to their gradual integration into the video domain. Clip4caption (Tang et al., 2021), Clip-DCD (Yang, Zhang, & Zou, 2022) and CroCaps (Xu et al.,

Fig. 2. Fine-grained statistics of scenes and entities in the VC-NBA-2022, Goal, and NSVA Datasets. The entity statistics refer to the average number of entities associated with each video.

2025) utilize CLIP for obtaining visual-text representations, markedly enhancing video captioning performance.

In this work, our EIKA employs the encoder–decoder framework and integrates the visual encoder from the pre-trained model CLIP4clip (Luo, Ji, Zhong, et al., 2022). Unlike most methods (Aafaq et al., 2019; Du, Zhu, Xiong, et al., 2023; Nabati & Behrad, 2023; Ye et al., 2022; Zeng, Wang, Liao, et al., 2024; Zhang et al., 2021) that use LSTM (Hochreiter & Schmidhuber, 1997) or GRU (Cho, Van Merriënboer, Gulcehre, et al., 2014) as the text decoder, we employ a transformer (Vaswani, Shazeer, Parmar, et al., 2017) structure as the decoder to effectively capture long-range visual and textual dependencies.

2.2. Sports video captioning

Sports video captioning is a challenging task because it involves describing multiple events including player-object (e.g., player and basketball) interactions and player-player interactions. Numerous works have been proposed across various sports domains. Yu, Cheng, Ni, et al. (2018) utilize a graph neural network to model the relationships among players, accurately mapping these complex interactions into detailed descriptions. Qi, Wang, Li, et al. (2019) design a model that generates descriptions of dynamic player/team movements and interactions in volleyball games, capturing player pose, trajectory, and group relationship features simultaneously. Zhang, Gao et al. (2024) propose a large basketball highlight commentary dataset and employ a simple yet effective real-time strategy to enhance multimodal feature interaction for generating emotional and descriptive commentary. The above methods focus on fine-grained actions but use macroscopic vocabularies (e.g., a man or a player) instead of player names, making it impossible to link players to the game scene.

Kim and Choi (2020) integrate information detected by various detectors with the domain ontology knowledge of baseball to generate descriptions that focus on player categories (e.g., batter) rather than their identities. Rao, Wu, Liu, Wang and Xie (2024) introduce the MatchTime dataset, derived from the anonymous soccer commentary dataset Soccer-Net Caption (Mkhallati et al., 2023), which exhibits strong alignment between textual and visual content. Subsequently, Rao, Wu, Jiang, Zhang and Xie (2024) propose the SoccerNetReplay-1988, the largest multimodal soccer dataset to date. In these soccer datasets, captions replace entities with special tokens like [PLAYER] and [TEAM], but all three methods fail to generate accurate role names. The above methods we mentioned consider entities, but still do not take into account the specific identities of entities, which leaves a certain distance from meeting practical applications.

Wu et al. (2022) integrate the visual information detected for players, basket, basketball, and the three-point line, generating accurate entity-aware captions. This method does not have any operation on entity recognition, but only uses the parameter memory learned from the training data. Qi et al. (2023) incorporate all pre-match and player-team information without a selection mechanism. This indiscriminate inclusion could introduce noise, affecting caption accuracy. Xi et al. (2025) introduce the extra game information (the candidate player list) to generate descriptions with player identities. These methods can extract player identity-related information from external explicit knowledge to generate entity-aware descriptions. However, explicit knowledge does not provide scene-related information, limiting the model's ability to distinguish different scenes and resulting in less accurate descriptions. In this work, we consider not only the generation of entity names but also the model's ability to distinguish different scenes, enabling the generation of higher-quality descriptions for sports videos.

2.3. Computer vision with knowledge

The introduction of knowledge enhances the model's cognitive understanding capability. Current approaches can typically be divided into two distinct types. The first type (Fang, Wang, Zhuo, et al., 2022; Gu et al., 2023; Li, Xu, Liu, et al., 2020; Zhuo, Zhu, Cui, et al.,

2022) focuses on enriching video features by extracting explicit knowledge from external knowledge bases. For example, to understand the causal relationships and social interactions underlying video content, works (Fang, Gokhale, Banerjee, et al., 2020; Shao, Fang, & Yang, 2022; Yu, Liang, Ji, et al., 2021) utilize the everyday commonsense reasoning database ATOMIC (Sap, Le Bras, Allaway, et al., 2019) to assist models in generating relevant commonsense descriptions. These methods acquire useful explicit knowledge from external knowledge base to help the model generate descriptions with fine-grained details. Inspired by this approach, to further generate fine-grained descriptions of sports videos, several methods (Qi et al., 2023; Xi et al., 2025) that incorporate game-related explicit knowledge from the knowledge graph into traditional video captioning models have brought video captioning a step closer to practical application.

The other type focuses on mining the implicit knowledge from model parameters (Ayyubi et al., 2023; Yuan, Jia, & Bao, 2023; Zhang et al., 2020) or databases (Huang, Wang, Zeng, & Wang, 2022; Ma et al., 2024; Zeng, Zhang, Gao, et al., 2023) to improve model performance. For example, Yuan et al. (2023) propose a knowledge guided network based on GPT-2 (Radford, Wu, Child, et al., 2019) for video-based commonsense captioning. This network leverages GPT-2 to enrich dataset knowledge, enabling models to learn commonsense not present in videos. To improve category detection accuracy, Du et al. (2025) propose a multimodal knowledge transfer method that incorporates cross-modal semantic information to learn the semantic relationships between categories. Zhou, Luo, and He (2025) propose a dynamic collaborative method based on heterogeneous knowledge transfer, where experts with different specializations work together to make predictions for long-tailed visual recognition tasks. Motivated by the fact that the human brain can well correlate arbitrary images with texts, Huang et al. (2022) calculate the average of image features associated with one conceptual word to obtain its visual semantic representation, dealing with unpaired image-text matching.

Contrastingly, inspired by the principle that humans retrieve previously formed scene concepts from memory and compare the new scene with these concepts to determine its category, our method utilizes the mean features of videos from different scenes as implicit scene knowledge which implies the definition and boundaries of the scene. It helps the model compare new video features with pre-defined scene knowledge to recognize different scenes. Furthermore, we utilize the multimodal information of players involved in the game as explicit knowledge to assist the model in generating descriptions with player names. The integration of explicit and implicit knowledge enhances the model's ability to recognize entities and scenes.

3. Method

3.1. Architecture overview

As shown in Fig. 3, the proposed EIKA mainly consists of 6 components: (1) Scene knowledge extraction (SKE) module. (2) Game knowledge extraction (GKE) module. (3) Spatial–temporal modeling module (STMM). (4) Video-knowledge interaction module (VKIM). (5) Entity-video interaction module (EVIM). (6) Scene-to-entity decoder (STE). We employ the CLIP (ViT-B/32) (Radford et al., 2021) as the visual encoder, which is pre-trained using large-scale video-text dataset in CLIP4clip. This well pre-trained visual encoder bridges the gap between video and text modalities, facilitating multimodal tasks such as video caption generation. The text encoder used in EIKA is identical to the one in CLIP4clip. STE and GKE provide implicit visual scene knowledge and explicit game knowledge for the model, respectively. VKIM utilizes the learnable query vectors to adaptively learn the scene-related features based on the scene knowledge and spatiotemporal video features. And EVIM combine the scene-related features and game knowledge to obtain the entity-related features, associating players with the video content. To fully utilize both scene and entity information, the STE decoder employs a scene-to-entity decoding strategy. First, the scene-related information is integrated by scene cross-attention, decoding scene-related textual features. Then, entity-related information is integrated through entity cross-attention, decoding entity-related textual features. In the following subsections, we will describe each component in the proposed EIKA in detail.

3.2. Scene Knowledge Extraction (SKE)

When viewing a video, people usually compare it with other videos that have similar scenes to obtain a more accurate understanding, and use similar videos to compensate for possible missing visual details in the source video. However, online retrieval of relevant information from external databases to enhance video features can increase computational costs. We explore the scene knowledge within the videos in a simple way. This implicit knowledge can not only provide complementary visual information, but also enhance the model's discrimination and understanding of visual scenes. Specifically, we utilize the visual encoder to extract the global features $V_{all}^{g} \in \mathbb{R}^{N_t \times 1 \times D_v}$ of all videos in training set, where D_v is the hidden size of visual encoder and N_t denotes the number of videos in the training set. We group the video features based on the labels annotated in the training set. The central features for various types of scenes are obtained by averaging the video features for each label. Formally,

$$C_r = \frac{1}{K} \sum_{v_j \in V_r^g} v_j, \tag{1}$$

where V_r^g is the set of video features with the label r. K is the number of features in this set, which corresponds to the number of videos with the label r. C_r denotes the central feature of label r.

Ultimately, the central features for R scenes are denoted as $C = \{C_1, C_2, \ldots, C_R\}$, $C \in \mathbb{R}^{R \times D_v}$. C is defined as the scene knowledge, which assists the model in distinguishing among different visual scenes and supplementing information deficiencies in the source video.

3.3. Game Knowledge Extraction (GKE)

In live sports broadcasting, commentators are typically provided with game-related information, such as the competing teams and the identities of each team's players. If the model has access to such explicit game-related knowledge, it would be better equipped to generate descriptions that include players' names.

In the basketball knowledge graph (Xi et al., 2025), each event (e.g., B. Ingram makes the 2-pt jump shot from 19 ft) is associated with a video clip and linked to a specific game. For each game, the players from both teams who participated are combined to form a candidate entity list, representing the game-related knowledge. Thus, each video clip has a corresponding player list. During the training or inference stage, the multimodal features of each player entity is obtained by combining their name features with their image features. Specifically, the global image features $E_p \in \mathbb{R}^{N_e \times D_v}$ and global name features $E_n \in \mathbb{R}^{N_e \times D_t}$ of all player entities in the candidate entity list are extracted by the visual encoder and text encoder, respectively. N_e is the number of entities in candidate entity list and D_t is the hidden size of text encoder. The image features and name features are added according to their corresponding positions to obtain the multimodal features $E_m \in \mathbb{R}^{N_e \times D_t}$.

$$E_m = E_p + E_n W_1, \tag{2}$$

where $W_1 \in \mathbb{R}^{D_v \times D_t}$ is the learnable matrix, which maps the visual feature to textual space.

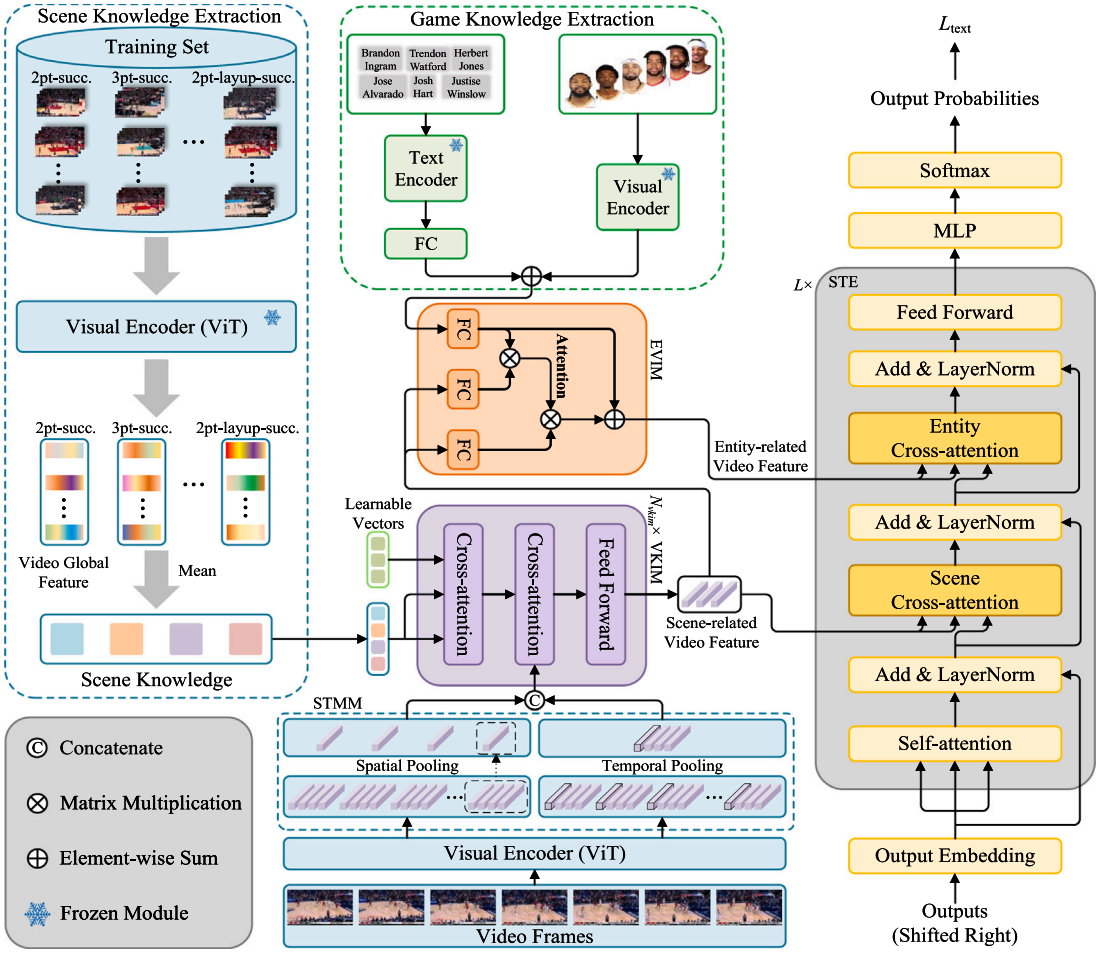


Fig. 3. Overview of the proposed Explicit & Implicit Knowledge-Augmented Network for Entity-Aware Sports Video Captioning (EIKA). The visual scene knowledge mined in the training set is utilized in both training and reasoning stage.

3.4. Spatial–temporal Modeling Module (STMM)

Videos contain not only spatial information from static frames but also dynamic behaviors and events that evolve over time, all of which are expressed through spatiotemporal features. Understanding these spatiotemporal dynamics is essential for accurately identifying objects, actions and the overall significance of scenes in videos. In this work, we design STMM to generate spatiotemporal features from a video clip $V \in \mathbb{R}^{T\times H\times W\times C}$ with T frames.

The visual encoder encodes T frames to visual embeddings $V_e \in \mathbb{R}^{T\times h\times w\times D_v}$, where $h = H/P$, $w = W/P$. P is the patch size (i.e., 32 for ViT-B/32). Visual embeddings V_e are then flatten to $V_f \in \mathbb{R}^{T\times N_p\times D_v}$, where $N_p = h\times w$. Visual embeddings V_f are averaged along the spatial dimension to obtain the spatial feature $V_s \in \mathbb{R}^{T\times D_v}$. Similarly, the visual embeddings V_f are averaged along the temporal dimension to obtain the temporal feature $V_t \in \mathbb{R}^{N_p\times D_v}$. The spatial feature and temporal feature are concatenated to yield the spatiotemporal feature V_{st} of the given video:

$$V_{st} = \left[V_s, V_t\right] \in \mathrm{R}^{(T+N_p)\times D_v}, \tag{3}$$

where $[,]$ denotes the concatenate function in Python.

3.5. Video-knowledge Interaction Module (VKIM)

The video-knowledge interaction module (VKIM) aims to learn the scene-related video feature $V_{scene} \in \mathbb{R}^{N_q\times D_v}$. As shown in Fig. 3, we randomly initialize N_q learnable query vectors $\Theta \in \mathbb{R}^{N_q\times D_v}$. Due to the abundance of redundant information in a video, directly interacting knowledge with the video would introduce a significant amount of noise features. Therefore, we adopt a learnable approach to adaptively capture key information. The learnable query vectors interact with scene knowledge $C \in \mathbb{R}^{R\times D_v}$ and spatiotemporal feature $V_{st} \in \mathbb{R}^{(T+N_p)\times D_v}$ in sequence to obtain scene-related video feature . We employ the multi-head cross-attention mechanism to perform the video-knowledge interaction as follow:

$$V_i' = \delta\left(\frac{\Theta W_i^{Q1}\cdot\left(CW_i^{K1}\right)^{\mathrm{T}}}{\sqrt{D_v}}\right)\cdot CW_i^{V1}, \tag{4}$$

$$V' = F_c^1\left(\left[V_1', V_2', \ldots, V_i'\right]\right), \tag{5}$$

$$V_i'' = \delta\left(\frac{V' W_i^{Q2}\cdot\left(V_{st}W_i^{K2}\right)^{\mathrm{T}}}{\sqrt{D_v}}\right)\cdot V_{st}W_i^{V2}, \tag{6}$$

$$V_i'' = FFN\left(F_c^2\left(\left[V_1'', V_2'', \ldots, V_i''\right]\right)\right), \tag{7}$$

where $W_i^{Q1} \in \mathbb{R}^{D_v\times D_v}$, $W_i^{K1} \in \mathbb{R}^{D_v\times D_v}$, $W_i^{V1} \in \mathbb{R}^{D_v\times D_v}$, $W_i^{Q2} \in \mathbb{R}^{D_v\times D_v}$, $W_i^{K2} \in \mathbb{R}^{D_v\times D_v}$ and $W_i^{V2} \in \mathbb{R}^{D_v\times D_v}$ are learnable matrices, i is the number of attention heads. $\delta(\cdot)$ denotes the softmax function. The outputs of all attention heads are integrated and fused through the fully-connected layer $F_c(\cdot)$. $FFN(\cdot)$ is the feed-forward network.

We stack N_{vkim} VKIM blocks to obtain a more refined scene-related video feature, and take the output of the last VKIM block as the final scene-related video feature V_{scene}.

3.6. Entity-video Interaction Module (EVIM)

The entity-video interaction module (EVIM) based on the attention mechanism can learn entity-related video feature $V_{entity} \in \mathbb{R}^{N_e \times D_v}$. We utilize the scaled dot-product attention function to associate the entities with a specific scene. The residual connection operation is utilized to enhance the feature representation.

$$V_{entity} = \delta\left(\frac{E_m W_2 \cdot \left(V_{scene} W_3\right)^{\mathrm{T}}}{\sqrt{D_t}}\right) \cdot V_{scene} W_4 + E_m W_2, \tag{8}$$

where $W_2 \in \mathbb{R}^{D_t \times D_v}$, $W_3 \in \mathbb{R}^{D_v \times D_v}$ and $W_4 \in \mathbb{R}^{D_v \times D_v}$ are learnable matrices.

3.7. Scene-to-entity (STE) decoder

We design the scene-to-entity decoder that can first decode scene-related textual features and then decode textual features with entity information, fully taking advantage of both scene and entity information. As shown in Fig. 3, scene cross-attention is employed to generate scene-related features $s^l_{1:t}$:

$$s^l_{1:t} = \ell\left(\mathcal{M}_{S-Att}\left(m^l_{1:t}, V_{scene}, V_{scene}\right) + m^l_{1:t}\right), \tag{9}$$

$$m^l_{1:t} = \ell\left(\mathcal{M}_{self}\left(f^{l-1}_{1:t}, f^{l-1}_{1:t}, f^{l-1}_{1:t}\right) + f^{l-1}_{1:t}\right), \tag{10}$$

where $\ell(\cdot)$ denotes the Layer Normalization and l denotes the current decoder layer. t is the decoding time and $m^l_{1:t}$ is the sentence feature of last self-attention module $\mathcal{M}_{self}(\cdot)$ layer's output. $\mathcal{M}_{S-Att}(\cdot)$ denotes the scene cross-attention module. The $f^0_{1:t}$ for the first layer is formulated as:

$$f^0_{1:t} = \tau_{pe}\left(\epsilon\left(w_{0:t-1}\right)\right), \tag{11}$$

where $\epsilon(\cdot)$ denotes the word embedding layer applied to the tokenized token w and $\tau_{pe}(\cdot)$ denotes the trigonometric positional embedding.

To generate entity-aware descriptions, the entity cross-attention is employed to incorporate entity information into textual features with scene context. The residual connection operation and the feed-forward network are employed to enhance the feature representation, and output the entity-related features $e^l_{1:t}$:

$$e^l_{1:t} = FFN\left(\ell\left(\mathcal{M}_{E-Att}\left(S^l_{1:t}, V_{entity}, V_{entity}\right) + S^l_{1:t}\right)\right), \tag{12}$$

where $\mathcal{M}_{E-Att}(\cdot)$ denotes the entity cross-attention module. We stack L decoder layer blocks to obtain the final the decoder's output e^L_t. The decoder combines probabilities and the word dictionary to decode the corresponding words. Output probabilities are obtained by MLP layer and softmax function as follow:

$$O_p = \delta\left(MLP\left(e^L_t\right)\right). \tag{13}$$

Following the standard training pattern for caption generation, we utilize cross-entropy loss to optimize our model:

$$\mathcal{L}_{\Phi} = -\sum_{t=1}^{N_g} \log\left(\mathrm{P}\left(w^*_t | w^*_{0:t-1}, V_{scene}, V_{entity}; \Phi\right)\right), \tag{14}$$

where $\left\{w^*_0, w^*_1, \ldots, w^*_{N_g}\right\}$ is the set of ground-truth tokenized tokens. N_g denotes the number of tokens. And Φ denotes the optimized parameters.

4. Experiments

To evaluate the performance of our designed model, EIKA is compared with the existing video captioning methods on three entity-aware sports video captioning datasets VC-NBA-2022 (Xi et al., 2025), Goal (Qi et al., 2023) and NSVA (Wu et al., 2022). We further carry out ablation studies to assess the impact and contribution of each individual component within EIKA.

4.1. Implementation details

The hidden size D_v of visual encoder is 768. And the hidden size D_t of text encoder is 512. The training epoch is set to 100 for both VC-NBA-2022 and Goal datasets. And the training epoch is set to 20 for NSVA dataset. Each video in VC-NBA-2022 and Goal datasets has T frames, which are sampled by using segment-based method (Wang, Xiong, Wang, et al., 2016). T is set to 18. Each frame size and entity picture size are 224 × 224. It is worth to note that the number N_e of candidate entities corresponding to each video is not fixed. We stack 4 video-knowledge interaction module blocks and 3 decoder layers for EIKA. And the number of learnable query vectors of video-knowledge interaction module is 18. Each of the learnable query vectors is randomly initialized. The number of heads in VKIM's multi-head cross attention is 8. During the training stage, the parameters of EIKA are optimized by BertAdam (Devlin, Chang, Lee, et al., 2018) with the learning rate of 3e–5 and weight decay of 1e–2. Moreover, the parameters of visual encoder and text encoder for candidate player list are frozen during training, and the visual encoder for video feature extraction is learnable. For reference stage, the beam size of beam search operation is set to 5.

4.2. Datasets

VC-NBA-2022 (Xi et al., 2025) is an entity-aware basketball video captioning dataset, which is collected from a NBA website and covers 25 games. It includes 9 types of fine-grained shooting events, knowledge of 286 players (i.e., images and names), and over 3.9k videos. The training set contains 3162 videos and the testing set contains 786 videos. Each video clip has one English description with entity names and a candidate player list. In this work, we utilize the candidate player list as the explicit knowledge and extract the implicit knowledge from the training set.

Goal (Qi et al., 2023) is a knowledge-grounded video captioning dataset for soccer commentary generation. It contains 20 full-game soccer videos, over 8.9k video clips, 22k sentences and 42k knowledge triples. Each video clip has one English description which is converted from commentator's audio speeches and the description is colloquial. In this work, we modify the format of the dataset to be the same as VC-NBA-2022. The training set has 5448 videos and the testing set has 661 videos. In addition, we save the names of teams and players in the dataset as the candidate entity list for providing explicit knowledge. When extracting implicit scene knowledge, we filter out several rare scenes, such as "Penalty", "Red card", and "Yellow->Red card". After adjustments, the training set contains 5434 videos.

NSVA (Wu et al., 2022) is a large-scale NBA dataset for sports video analysis, which is built on web data and covers 132 games. This dataset consists of 32 019 video clips for fine-grained video captioning, action recognition and player identification. For captioning, it contains 23 804 training videos and 768 testing videos. The caption contains the distance between the ball and the basket, which brings great challenges to the generation performance of the model. Although this dataset is entity-aware, the player names in the descriptions appear in the form of ID number, such as "Player1629028". In this work, we collect the players involved in each game to form a candidate player list, serving as explicit knowledge. When extracting implicit scene knowledge, we filter out several rare scenes, such as "Ejection" and "Violation". After adjustments, the training set contains 23 790 videos.

We provide word-cloud based statistics in Fig. 4 to highlight the thematic focus and coverage of each dataset. Dataset VC-NBA-2022 (Fig. 4(a)) primarily focuses on various fine-grained shot and rebound events in basketball and the top-4 words in this dataset are "jump", "shot", "3pt" and "defensive". Dataset Goal (Fig. 4(b)) is dedicated to real-time commentary in soccer games, containing a substantial amount of background knowledge and colloquial expressions. Compared to VC-NBA-2022, dataset NSVA (Fig. 4(c)) covers a broader range of basketball events (e.g., foul and turnover) and is more extensive in size.

(a) VC-NBA-2022 (b) Goal (c) NSVA

Fig. 4. Word cloud of VC-NBA-2022, Goal and NSVA datasets. The bigger the font, the more percentage it occupies.

Table 1
Quantitative comparison results on VC-NBA-2022 dataset. * denotes that the model is equipped with explicit game-related knowledge. ** denotes that the model is equipped with explicit game-related knowledge and implicit scene knowledge. Note that KEA* denotes that KEA is equipped with implicit scene knowledge. Numbers in bold denote the best performance.

Model	Year	CIDEr	METEOR	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
V2C	2020	13.7	18.7	45.9	50.1	39.5	27.0	14.9
C4C	2021	70.4	26.7	51.2	49.1	42.5	35.4	28.8
C4C*	2025	117.2	27.5	52.0	50.5	43.7	37.0	29.4
C4C**	2025	124.2	27.8	52.8	51.0	43.9	37.4	30.2
SwB	2022	69.1	26.5	49.0	47.8	41.4	34.5	28.4
SwB*	2025	120.1	27.9	52.1	51.3	43.9	37.4	30.3
SwB**	2025	128.0	28.3	52.8	51.7	44.2	37.8	30.9
CCP	2023	70.5	27.4	50.7	49.8	42.3	35.6	28.9
CCP*	2025	122.3	28.3	52.9	52.2	44.6	38.0	30.3
CCP**	2025	129.5	28.8	53.4	52.9	45.0	38.4	31.6
OVD	2024	71.2	27.5	50.7	49.7	42.4	36.5	29.2
OVD*	2025	125.2	28.6	53.3	52.5	45.2	38.7	30.5
OVD**	2025	130.8	29.0	56.0	53.7	46.6	39.7	32.6
KEA	2025	138.5	28.0	54.9	53.1	46.4	38.8	32.4
KEA*	2025	**144.6**	28.7	56.1	54.5	47.6	40.2	34.4
Ours	2025	140.7	**29.5**	**56.8**	**57.1**	**50.3**	**43.1**	**36.7**

4.3. Evaluation metrics

We employ the standard captioning evaluation metrics, including BLEU (B) (Papineni, Roukos, Ward, et al., 2002), Rouge-L (R) (Lin, 2004), METEOR (M) (Banerjee & Lavie, 2005) and CIDEr (C) (Vedantam, Lawrence Zitnick, & Parikh, 2015). BLEU primarily evaluates text quality by comparing the overlap of n-grams between generated text and reference text. Rouge-L focuses on measuring the overlap between automatically generated text and reference text using the longest common subsequence method. METEOR combines precision, recall, and word order in its evaluation, closer to human evaluation. CIDEr evaluates the quality of generated text by considering the uniqueness of words and the consensus of sentences, reflecting the informativeness and accuracy of the text. Within these evaluation metrics, CIDEr is specially designed for captioning task evaluation, whereas BLEU, ROUGEL and METEOR are adapted from conventional machine translation. Following Qi et al. (2023), we only report the comparative results for metrics CIDEr, METEOR, Rouge-L and BLEU-1 on Goal dataset.

4.4. Compared methods

In this section, we provide the brief descriptions of the methods compared in our experiments:

- **V2C** (Fang et al., 2020): V2C utilizes a CNN-based video encoder to extract global features of the given video. A cross-modal self-attention module is employed to encode the joint visual-textual features and a transformer decoder is employed to generate captions.
- **Clip4Caption (C4C)** (Tang et al., 2021): C4C enhances video captioning by utilizing a CLIP-augmented video-text alignment network and employs a transformer decoder to generate captions.
- **SwinBERT (SwB)** (Lin et al., 2022): SwB is an end-to-end transformer-based video captioning network. It utilizes VidSwin (Liu et al., 2022) to extract spatiotemporal features from video and generate captions through a sequence-to-sequence mechanism.
- **CoCap (CCP)** (Shen et al., 2023): CoCap combines the motion vector, residual, and video features through the action encoder. The fused features are then passed to the decoder to generate video captions.
- **OmniViD (OVD)** (Wang et al., 2024): OmniViD is a unified generative framework designed to address various video tasks, including action recognition, captioning, video question answering, dense video captioning, and visual object tracking.
- **KEANet (KEA)** (Xi et al., 2025): KEA utilizes the candidate player list as the game-related knowledge to generate captions with player names.
- **MP-LSTM** (Venugopalan et al., 2014): MP-LSTM utilizes LSTM (Hochreiter & Schmidhuber, 1997) to generate captions based on the video features.
- **TA** (Yao et al., 2015): TA improves video captioning by introducing a temporal attention module to exploit global temporal structure, enhancing the correlation between each frame and the words generated at the previous moment.
- **T2T** (Sharma, Ding, Goodman, & Soricut, 2018): T2T utilizes the transformer-based structure to achieve the conceptual captioning.
- **UniVL** (Luo et al., 2020): UniVL is a unified video-language pre-training model that employs an encoder–decoder architecture, featuring separate unimodal encoders and a cross-modal encoder.

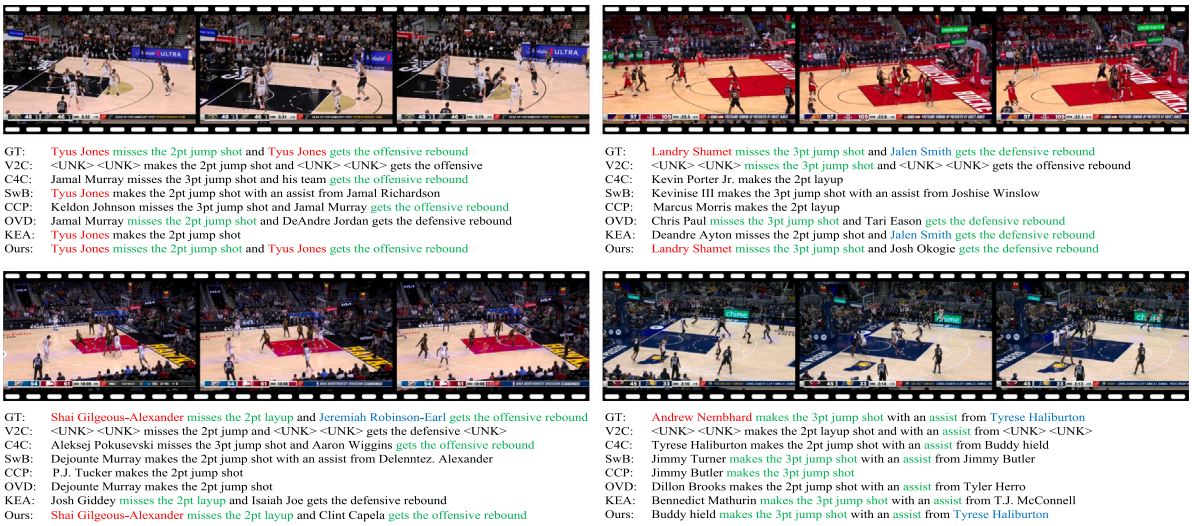


Fig. 5. Qualitative comparison results of our model and five video captioning models on VC-NBA-2022 dataset. GT, V2C, C4C, SwB, CCP, OVD and KEA denote ground truth caption, Video2Commonsense, Clip4Caption, SwinBERT, CoCap, OmniViD and KEANet, respectively. Different entity names are marked red and blue. And the specific visual scenes are marked green.

Table 2
Quantitative comparison results on Goal dataset. * denotes that the model is equipped with explicit game-related knowledge. ** denotes that the model is equipped with explicit game-related knowledge and implicit scene knowledge. Note that KEA* denotes that KEA is equipped with implicit scene knowledge. Numbers in bold denote the best performance.

Model	Year	CIDEr	METEOR	Rouge-L	BLEU-1
V2C	2020	0.1	2.1	3.4	3.4
C4C	2021	2.2	5.0	5.5	5.7
C4C*	2025	2.5	5.2	5.8	5.9
C4C**	2025	2.7	5.5	6.0	6.2
SwB	2022	2.2	5.1	5.3	5.7
SwB*	2025	2.6	5.5	6.0	5.9
SwB**	2025	2.7	5.7	6.1	6.1
CCP	2023	2.3	5.0	5.3	5.5
CCP*	2025	2.5	5.3	5.8	5.9
CCP**	2025	2.7	5.9	6.0	6.2
OVD	2024	3.0	5.9	9.1	10.7
OVD*	2025	3.9	6.4	10.6	14.4
OVD**	2025	**4.2**	6.5	11.0	**15.4**
KEA	2025	3.7	6.4	10.5	14.9
KEA*	2025	4.0	**6.6**	10.8	15.2
Ours	2025	4.1	**6.6**	**11.2**	15.3

It is trained to learn joint video-text representations through various objectives, such as video-text alignment and masked modeling, enabling strong performance across multimodal tasks like video captioning and video retrieval.

- **NSVA** (Wu et al., 2022): NSVA improves upon UniVL by jointly utilizing video, basketball, basket, player, and court features to perform player identity-aware video captioning.

Since the compared methods are traditional video captioning models not specifically designed for entity-aware sports video captioning, we equip each with explicit and implicit knowledge to transform them into entity-aware models for a fair comparison.

4.5. Performance comparison

Comparison on VC-NBA-2022. Our EIKA model is compared with V2C (Fang et al., 2020), C4C (Tang et al., 2021), SwB (Lin et al., 2022), CCP (Shen et al., 2023), OVD (Wang et al., 2024) and KEA (Xi et al., 2025) on VC-NBA-2022 dataset. V2C, C4C, SwB and OVD take only videos as input. CCP takes the motion vector, residual, and video features as input. KEA takes the explicit game-related knowledge and video as input. As shown in Table 1, compared with the baseline C4C, EIKA gains 70.3% and 7.9% absolute improvements on the metrics CIDEr and BLEU-4, respectively. EIKA outperforms CCP by 70.2% and 7.8% on CIDEr and BLEU-4, respectively. And EIKA outperforms OVD by 70.2% and 7.5% on CIDEr and BLEU-4, respectively.

Compared to KEA, our EIKA surpasses the single explicit knowledge-guided model on all metrics. This validates that EIKA can generate more accurate captions with entity names. This is because we introduce scene knowledge to guide the model in distinguishing different visual scenes, generating accurate scene categories. EIKA gradually generates detailed descriptions, moving from scenes to entities. With the addition of implicit scene knowledge, the KEA* model shows improvement across all metrics. It achieves a higher CIDEr score than our EIKA. This is because KEA* employs a more complex T5 large language model (Raffel et al., 2020) as its decoder, while our model only utilizes a 3-layer Transformer as the decoder. The T5 model significantly enhances KEA*'s generation capabilities, particularly in generating fine-grained information such as player names. Additionally, KEA* processes video frames at the resolution of 1280 × 720, much higher than our input of 224 × 224, enabling more accurate visual feature extraction and scene understanding. The CIDEr metric, designed for captioning tasks, is particularly sensitive to rare terms (such as proper nouns and infrequent words), assigning especially high weight to player names. Given KEA*'s higher-resolution input and more powerful decoder, its video understanding capability is further enhanced after integrating implicit scene knowledge, resulting in more pronounced CIDEr score differences on this task. In contrast, other metrics (e.g., METEOR, ROUGE-L, BLEU) focus more on n-gram matching and semantic alignment, with less dependency on rare terms, indicating that our generated captions maintain high syntactic and semantic similarity. Notably, KEA* achieves a higher CIDEr score only after integrating our proposed implicit scene knowledge. This indirectly confirms the positive impact of implicit scene knowledge on description generation, further validating its effectiveness. Moreover, when the model is equipped with both explicit and implicit knowledge, its performance is further improved. For example, compared to OVD, OVD** has improved by 59.6% and 3.4% on evaluation metrics CIDEr and BLEU-4, respectively. This further validates the effectiveness of combining the two types of knowledge.

Fig. 5 shows the qualitative results on VC-NBA-2022 dataset. Since V2C does not have any tokenizer tools, it cannot encode and decode

Fig. 6. Qualitative comparison results of our model and five video captioning models on Goal dataset. GT, V2C, C4C, SwB, CCP, OVD and KEA denote ground truth caption, Video2Commonsense, Clip4Caption, SwinBERT, CoCap, OmniViD and KEANet, respectively. Different entity names are marked red and blue. And the specific visual scenes are marked green.

Table 3
Quantitative comparison results on NSVA dataset. S3D and T denote the S3D feature and Timesformer feature, respectively. "(full)" denotes the method jointly utilize video, basketball, basket, player, and court features. "list" and "scene" denote the explicit knowledge and implicit knowledge, respectively. NSVA* is the improved method that incorporates knowledge.

Model	Feature	Year	CIDEr	METEOR	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MP-LSTM	S3D	2014	50.0	15.3	33.2	32.5	23.6	16.7	12.1
TA	S3D	2015	54.6	15.6	34.0	33.1	24.2	17.5	12.8
T2T	S3D	2018	57.2	16.1	35.7	34.6	25.4	18.1	13.1
UniVL	S3D	2020	71.7	19.2	40.1	44.1	30.9	22.6	16.9
NSVA	S3D(full)	2022	98.6	22.7	46.6	47.9	37.1	28.1	21.6
	T	2022	95.6	21.7	46.8	46.7	36.3	27.4	20.9
	T(full)	2022	113.9	24.3	50.8	52.2	41.0	31.4	24.3
NSVA*	S3D+list	2025	75.3	20.4	43.1	45.7	32.2	24.5	17.9
	S3D+list+scene	2025	78.2	22.5	46.5	46.9	35.6	27.3	21.2
	T+list	2025	97.2	23.1	48.7	49.5	37.4	29.2	21.8
	T+list+scene	2025	99.7	24.6	50.3	50.9	39.5	31.8	23.1
	S3D(full)+list	2025	99.8	23.5	47.2	48.3	37.9	29.2	22.3
	S3D(full)+list+scene	2025	101.5	24.2	48.4	48.8	38.6	30.3	23.4
	T(full)+list	2025	115.2	24.8	51.7	52.8	41.8	32.5	25.3
	T(full)+list+scene	2025	117.4	25.6	52.7	53.5	42.4	33.5	26.1

Table 4
Ablation study on VC-NBA-2022 dataset. GK, EVIM, SK, VKIM and STMM denote game knowledge and entity-video interaction module, scene knowledge, video-knowledge interaction module and spatial–temporal modeling module, respectively. Numbers in bold denote the best performance.

Model	GK	EVIM	SK	VKIM	STMM	CIDEr	METEOR	Rouge-L	BLEU-4
A						70.4	26.7	51.2	28.8
B	✓					117.2	27.5	52.0	29.4
C	✓	✓				125.1	28.0	53.2	31.6
D			✓			88.7	26.9	51.4	29.0
E			✓	✓		92.2	27.2	51.7	30.5
F					✓	75.6	27.1	51.6	29.1
G	✓	✓			✓	127.4	28.5	54.0	32.0
H			✓	✓	✓	95.2	27.4	52.7	31.0
I	✓	✓	✓		✓	137.1	29.0	55.3	33.8
J	✓		✓	✓	✓	135.0	28.8	55.4	33.9
K	✓	✓	✓	✓	✓	**140.7**	**29.5**	**56.8**	**36.7**

entity names. So entity names are directly replaced with the special token <UNK>. C4C, SwB, CCP and OVD struggle to accurately generate entity names and fine-grained scene categories. In contrast, under the guidance of knowledge, the generation performance of KEA and EIKA is improved. In the comparison between KEA and EIKA, it can be seen that EIKA can recognize more accurate visual scenes. This benefits from the guiding role of scene knowledge. The above results indicate that EIKA is more closely aligned with practical applications.

Comparison on Goal. EIKA is compared with V2C, C4C, SwB, CCP, OVD and KEA on Goal dataset. As shown in Table 2, EIKA outperforms the baseline C4C by 1.9% and 9.6% on CIDEr and BLEU-1, respectively. EIKA outperforms the previous state-of-the-art model KEA on all metrics. When the model combines explicit and implicit knowledge, its performance is further improved. Notably, the decoder of OVD is the BART large langugae model (Lewis, 2019), which excels in generating long and coherent text with accurate contextual dependency.

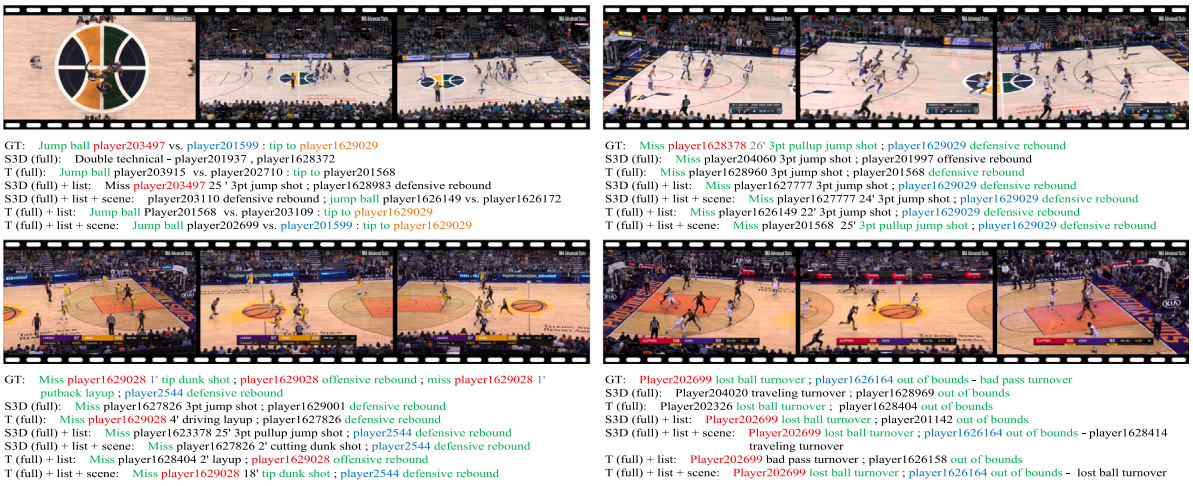


Fig. 7. Qualitative comparison results with different knowledge for different model settings on NSVA dataset. Different entity names are marked red, blue and orange. And the specific visual scenes are marked green.

When equipped with implicit scene knowledge, OVD** demonstrates a stronger understanding of video content, producing long-form captions with more accurate player names and scene types. Consequently, OVD** achieves higher CIDEr and BLEU-1 scores than our EIKA. Additionally, after integrating implicit scene knowledge, KEA* achieves METEOR scores comparable to ours. Overall, both KEA* and OVD** exhibit performance improvements, validating the effectiveness of jointly utilizing both types of knowledge.

Fig. 6 shows the qualitative results on Goal dataset. Both KEA and EIKA can generate partial correct entity names and scenes. However, EIKA can identify more accurate visual scenes. The results above indicate that, with the guidance of explicit game-related and implicit scene knowledge, the model can identify entity names and specific scenes in the challenging task of sports commentary.

Comparison on NSVA. Since NSVA does not provide raw video data but instead offers features extracted using Timesformer (Bertasius, Wang, & Torresani, 2021), we incorporate both explicit and implicit knowledge as additional modules to UniVL (Luo et al., 2020) and NSVA (Wu et al., 2022) for a fair comparison. This approach explores the effectiveness of the two types of knowledge across different models.

As shown in Table 3, models in the NSVA* category exhibit performance improvements over UniVL and NSVA when combined with different types of knowledge. For example, NSVA*(S3D + list) and NSVA*(S3D + list + scene) improve the CIDEr score of UniVL by 3.6% and 6.5%, respectively. NSVA*(T(full) + list) and NSVA*(T(full) + list + scene) improve the CIDEr score of NSVA(T(full)) by 1.3% and 3.5%, respectively.

Since the NSVA dataset is constructed with the principle that each team's game appears at least once in the training set (Wu et al., 2022), the model can leverage parameter memory to predict player identities during testing. As a result, the performance gains from adding the list knowledge (explicit game knowledge) are relatively small. Additionally, the dataset's captions include references to the distance between shooting positions and the basket, presenting a challenge for the generation task. However, the introduction of multiple types of knowledge still yields noticeable performance improvements. As illustrated in Fig. 7, incorporating the list knowledge (explicit game knowledge) leads to more accurate predictions of player identities within the text. Furthermore, adding scene knowledge (implicit scene knowledge) makes scene category generation more accurate.

4.6. Ablation study

Effectiveness of each component. We conduct a series of ablation experiments on VC-NBA-2022 with the metrics of CIDEr, METEOR, Rouge-L and BLEU-4. Table 4 shows the results of the ablation experiments. Model A is the baseline Clip4Caption. Model B introduces explicit game-related knowledge on the basis of Model A, greatly improving performance. Model A can generate descriptions with entity names based solely on the memory of trained parameters. However, when the player entities in the video do not appear in the training data, model A predicts incorrectly, and the names might belong to other games. By utilizing game knowledge, the model can obtain entities information from a limited range and improve the accuracy of names prediction. Model C introduces entity-video interaction module (EVIM) on the basis of Model B, helping the model focus on video-related entities. Model D introduces scene knowledge into Model A, leading 18.3% and 0.2% improvement in CIDEr and BLEU-4 scores, respectively. This is attributed to scene knowledge, which contains the definitions of various visual scenes and the boundaries among scenes. With the help of video-knowledge interaction module (VKIM), Model E adaptively captures knowledge-related video information. The CIDEr score is increased by 3.5% and the BLEU-4 score is increased by 0.5%. From the comparison results of (H, E), (G, C), and (F, A), it can be found that better visual features are beneficial for the model to recognize specific scenes and entities. This is because spatial–temporal modeling module (STMM) can capture dynamics and contextual information in video, improving the model's understanding of video content. The comparison results of (G, I) and (H, J) validate the effectiveness of the two types of knowledge. Model K achieves optimal performance by adding all components to Model A. Model K outperforms Model A by 70.3%, 2.8%, 5.6% and 7.9% on the metrics of CIDEr, METEOR, Rouge-L and BLEU-4, respectively. The above results demonstrate the effectiveness of all components.

Table 5
Comparison results of model performance guided by different types of decoder on VC-NBA-2022 dataset. Numbers in bold denote the best performance.

Model	E	SCE	ETS	STE	CIDEr	METEOR	Rouge-L	BLEU-4
#1	✓				125.5	27.6	53.4	31.1
#2		✓			131.2	28.5	54.2	31.8
#3			✓		138.4	28.8	55.1	32.2
#4				✓	**140.7**	**29.5**	**56.8**	**36.7**

Impact of different decoding methods. As shown in Fig. 8, there are 4 types of decoding methods. The first method only sends entity-related video feature V_{entity} into the decoder. The second method is to concatenate scene-related video features and entity-related video features (Scene-concatenate-entity, SCE), and then feed them into the decoder. The third method is to first feed entity-related features into the

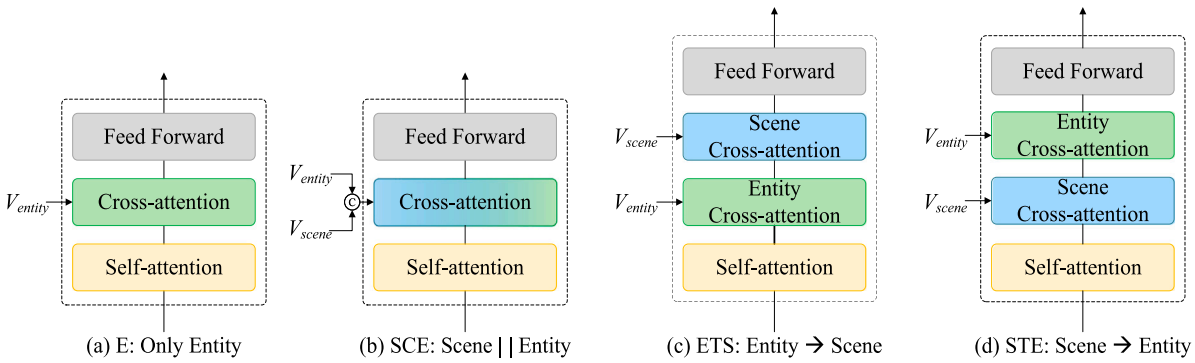


Fig. 8. Decoder variants for different decoding methods.

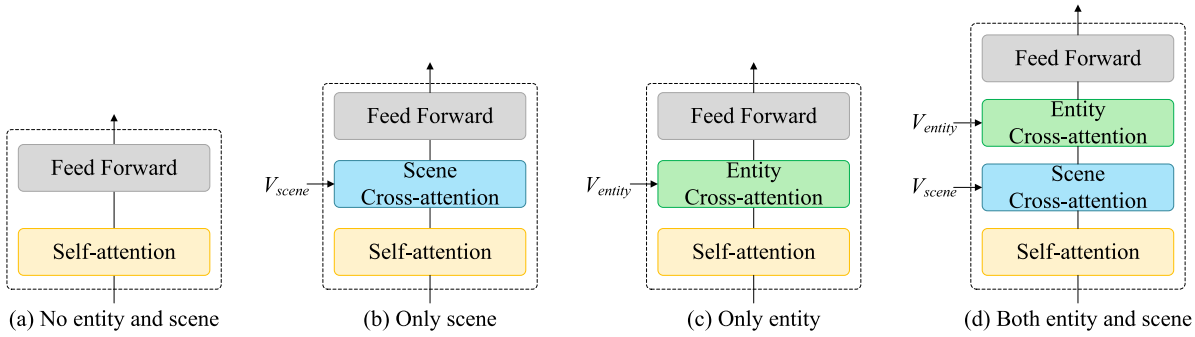


Fig. 9. Decoder variants for different cross-attention configurations.

decoder, and then feed scene-related features into the decoder (Entity-to-scene, ETS). The fourth method is opposite to the second method (Scene-to-entity, STE).

As shown in Table 5, the model has the lowest performance with only entity-related video feature (#1). And it can be seen that the performance of distributed decoding is better than that of cascading. The second cascading feature (#2) decoding method may lead to a decrease in text quality due to potential interference between feature information. Directly concatenating scene-related video features and entity-related video features may hinder effective information fusion, making it challenging for the decoder to understand the complex relationships between explicit external information and visual features, as well as decode multiple fine-grained details simultaneously. To alleviate these issues, the STE decoder employs a scene-first, entity-later decoding strategy. First, the scene-related information is integrated by scene cross-attention, decoding scene-related textual features. Then, entity-related information is integrated through entity cross-attention, decoding entity-related textual features. This approach fully leverages both scene and entity information. The step-by-step decoding strategy alleviates the difficulty of decoding multiple fine-grained details at once, improving the accuracy of text generation. Compared with the third method (#3), the fourth method (#4) of defining the scene first and then adding entity features can better utilize the contextual information provided by the scene for entity behavior and generate more logical and coherent descriptions.

Different cross-attention configurations of decoder. As shown in Fig. 9, there are 4 types of cross-attention configurations in the decoder. The first type does not utilize either scene cross-attention or entity cross-attention. The second type utilizes only scene cross-attention, while the third type utilizes only entity cross-attention. The fourth type has both scene cross-attention and entity cross-attention.

As shown in Table 6, when neither scene nor entity cross-attention is applied, the decoder relies solely on video content without incorporating any scene- or entity-related information, resulting in the poorest performance. When the decoder is equipped with either scene or entity cross-attention, it integrates scene-level contextual information or entity-specific semantic features, leading to performance improvements. The CIDEr score increases from 75.6% to 95.2% and 127.4%, respectively. And these results indicate that entity-related information contributes more significantly to performance gains. When both scene and entity cross-attention layers are applied simultaneously, the model achieves the best performance, with CIDEr and BLEU-4 scores improving by 65.1% and 7.6%, respectively. This further validates the effectiveness of jointly leveraging both explicit and implicit knowledge.

Impact of different approaches for obtaining implicit scene knowledge. We investigate the impact of 3 different approaches for obtaining scene knowledge on model performance. The first approach (A1) is the K-means algorithm. It directly clusters the training video features to 9 centers. The second approach (A2) is also a clustering algorithm. It clusters the video features in each category to yield one center. The third approach (A3) averages the features of videos for each category. As shown in Table 7, model ① does not use any scene knowledge. Model ② utilizes the scene knowledge obtained by the first approach, but its performance declines. This is because live sports videos contain many visually similar scenes, and the K-means algorithm incorrectly treats visual features from different scenes as one category. Both the second and third approaches involve mining scene knowledge of certain scenes within the same label, which avoids the issue of visual features confusion. Model ④ performs better than model ③. This is because the third approach of mining scene knowledge directly reflects the average features of each category and can better capture the commonalities within categories. However, the second approach relies on the operation of the K-means algorithm, which can be affected by the choice of initial centers, internal category variability, and other factors, resulting in the final cluster centers not necessarily being the optimal representatives of the categories. Therefore, we choose the scene knowledge obtained by the third approach to guide the model in recognizing different visual scenes.

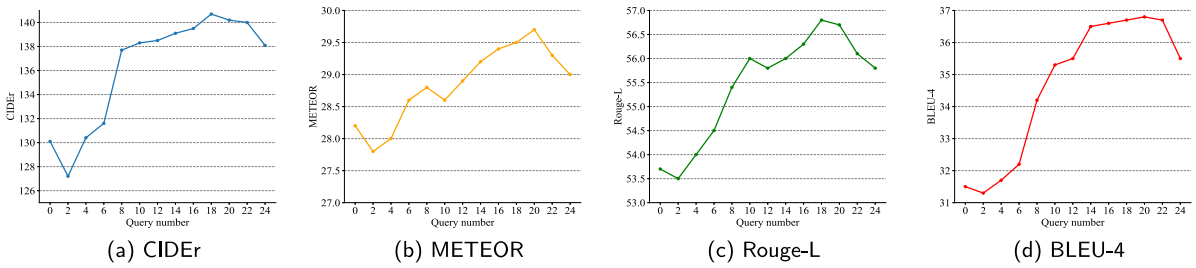


Fig. 10. Impact of the hyper-parameters in VKIM. Ablation studies of different query vector number for VKIM.

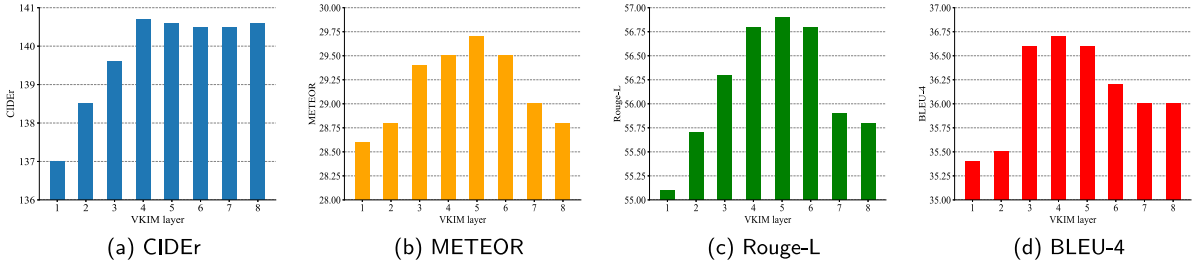


Fig. 11. Impact of the hyper-parameters in VKIM. Ablation studies of different layer number for VKIM.

Table 6
Comparison results of model performance with different cross-attention configurations of decoder layer on VC-NBA-2022 dataset. Numbers in bold denote the best performance.

Model	Scene	Entity	C	M	R	B-4
a			75.6	27.1	51.6	29.1
b	✓		95.2	27.4	52.7	31.0
c		✓	127.4	28.5	54.0	32.0
d	✓	✓	**140.7**	**29.5**	**56.8**	**36.7**

Table 7
Comparison results of model performance guided by scene knowledge obtained by different approaches on VC-NBA-2022 dataset. Numbers in bold denote the best performance.

Model	A1	A2	A3	C	M	R	B-4
①				127.4	28.5	54.0	32.0
②	✓			120.3	27.3	52.4	30.1
③		✓		138.9	28.9	55.2	36.2
④			✓	**140.7**	**29.5**	**56.8**	**36.7**

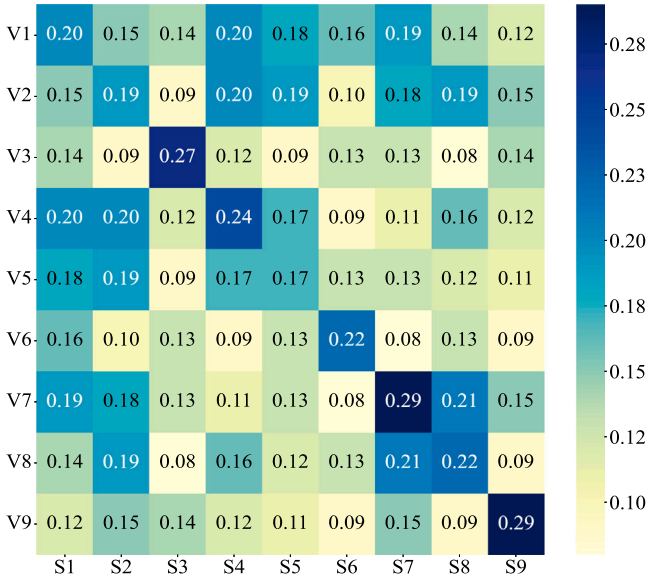


Fig. 12. Visualization of average similarity scores between scene knowledge and video features. The scene labels from S1 to S9 represent "2p-succ.", "2p-fail-off.", "2p-fail-def.", "2p-layup-succ.", "2p-layup-fail-off.", "2p-layup-fail-def.", "3p-succ.", "3p-fail-off.", and "3p-fail-def". And the video labels from V1 to V9 represent "2p-succ.", "2p-fail-off.", "2p-fail-def.", "2p-layup-succ.", "2p-layup-fail-off.", "2p-layup-fail-def.", "3p-succ.", "3p-fail-off.", and "3p-fail-def". The x-axis represents the categories of scene knowledge, and the y-axis represents the labels of different video categories.

Impact of the hyper-parameters in VKIM. We exploit how the number of query vector in VKIM affects the performance of EIKA. As shown in Fig. 10, when the number is set to 18, the model performs the best. When the number is less than 18 or greater than 18, the performance of the model declines. A smaller number results in capturing less knowledge-related video information, while a larger number leads to the introduction of redundant video information. The number of query vectors for VKIM is ultimately set to 18. Different layers of VKIM also affect the performance of the model. As shown in Fig. 11, when the number of layers is 4, the performance of the model is optimal. Excessive layers not only increase the computational complexity of the model, but also make it difficult to fit the model to the optimal performance during training. Therefore, the number of layers for VKIM is set to 4.

Qualitative analysis on scene knowledge. To explore the similarity between scene knowledge and all videos, we evaluate the similarity between the scene knowledge of each category and videos from different categories. The similarity scores for different groups (e.g., S1-V1, S1-V8 and so on) are then averaged. As shown in Fig. 12, most scene knowledge demonstrates the highest similarity with video features in the same categories, indicating that this knowledge effectively represents the scene concepts of those video categories. However, the similarity between scene knowledge S2 ("2p-fail-off.") and videos of the same category is lower than that with V4 ("2pt-layup-succ."). The similarity between scene knowledge S5 ("2p-layup-fail-off.") and videos of the same category is lower than that with V1 ("2pt-succ.") and V2 ("2p-fail-off."). One possible reason is the limited number of videos for these two scenes in the dataset, resulting in less accurate conceptual knowledge. Another reason is that these labels have similar scenarios, such as all having a two-point shot and all being within

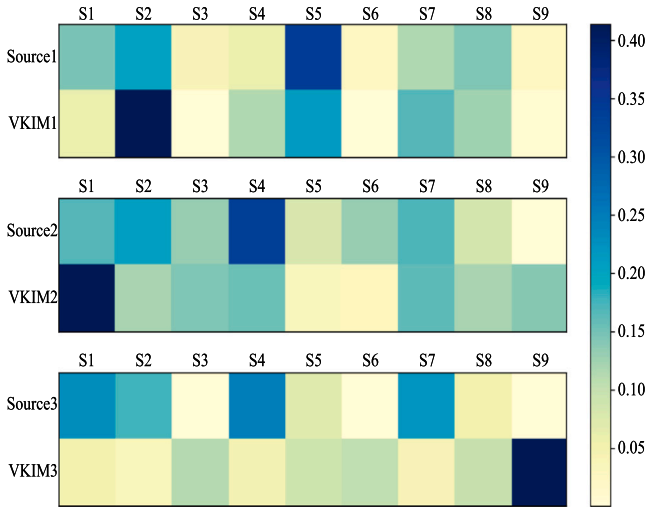


Fig. 13. Visualization of the similarity score between scene knowledge and two kinds of video features.

the three-point line. These visualization results demonstrate that our scene knowledge not only connects with videos corresponding to the same scene but also distinguishes features of most different videos. This aligns with the principle of category-based reasoning in human cognitive psychology. Our scene knowledge serves as conceptual scene representations, enabling the model to compare new concepts more effectively and thereby achieve more accurate scene recognition.

To intuitively evaluate the effectiveness of scene knowledge, we visualize the similarity scores between scene knowledge with 9 center features and two kinds of video features (source spatiotemporal video feature and scene-related video feature). As shown in Fig. 13, we present 3 groups of comparison results. S1-S9 are center features of 9 specific scenes. The ground truth scenes for the 3 groups of videos are S2, S1, and S9, respectively. Taking the first group as an example, the spatiotemporal feature of source video shows the highest similarity with scene S5. Under the guidance of visual scene knowledge, the scene-related video features learned by VKIM exhibit the highest similarity with scene S2. These 3 groups of visualization comparison results demonstrate that scene knowledge can assist the model in distinguishing different visual scenes, thereby enabling the generation of more accurate visual scene categories.

We further provide the t-SNE (Van der Maaten & Hinton, 2008) visualization results of source spatiotemporal video feature and scene-related video feature (Fig. 14). Under the guidance of scene knowledge, the categories of each scene are easier to distinguish. This further confirms that mining implicit knowledge within videos helps the model distinguish different scenes.

5. Conclusion, limitation and future work

In this paper, we mine the implicit visual scene knowledge from the training set by a simple way. And we propose an Explicit & Implicit Knowledge-Augmented Network for Entity-Aware Sports Video Captioning (EIKA), which effectively incorporates explicit game knowledge and implicit scene knowledge to generate descriptions with specific entity names. Our proposed EIKA achieves advanced performance on multiple entity-aware datasets. Extensive ablation studies and qualitative results have demonstrated the effectiveness of two types of knowledge and each component in EIKA.

Although our method performs well on dataset with neatly formatted descriptions (e.g., VC-NBA-2022), it shows lower performance on dataset with more casual and random text styles (e.g., Goal). This indicates that the knowledge we currently utilize is insufficient. In

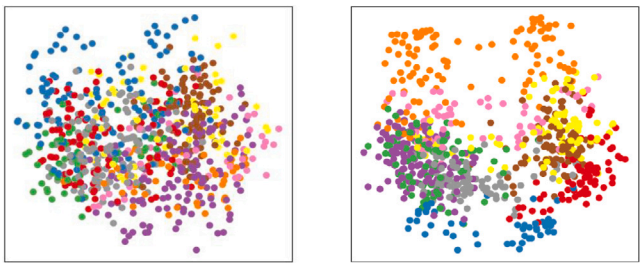


Fig. 14. t-SNE (Van der Maaten & Hinton, 2008) visualization results of source spatiotemporal video feature and scene-related video feature.

the future, we will try to acquire more comprehensive and informative knowledge from sports-related knowledge graphs and leverage the powerful generative capabilities of large models to achieve more complex sports video captioning tasks. Furthermore, while using the average features serves as an approximate representation of knowledge, its limitation lies in ignoring higher-order statistical information within the data distribution. Future work could incorporate variance or other distributional features to more comprehensively characterize scene knowledge, thereby enhancing the accuracy and robustness of the method.

CRediT authorship contribution statement

Zeyu Xi: Conceptualization, Methodology, Software, Investigation, Writing – review & editing, Visualization, Data curation. **Ge Shi:** Conceptualization, Writing – review. **Haoying Sun:** Data curation, Formal analysis. **Bowen Zhang:** Data curation, Validation. **Shuyi Li:** Conceptualization, Formal analysis. **Lifang Wu:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

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.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62236010, 62106010, 62306021; in part by the Beijing Natural Science Foundation under grant L233008.

Data availability

Data will be made available on request.

References

Aafaq, N., Akhtar, N., Liu, W., et al. (2019). Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 12487–12496).

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

Banerjee, S., & Lavie, A. (2005). 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, (pp. 65–72).

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *vol. 2*, In *ICML* (p. 4).

Chen, S., & Jiang, Y.-G. (2021). Motion guided region message passing for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* ICCV, (pp. 1543–1552).

Cho, K., Van Merriënboer, B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Cook, A., & Karakuş, O. (2024). LLM-Commentator: Novel fine-tuning strategies of large language models for automatic commentary generation using football event data. *Knowledge-Based Systems*, *300*, Article 112219.

Devlin, J., Chang, M.-W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Du, Y., Liu, F., Jiao, L., Li, S., Hao, Z., Li, P., et al. (2025). Text generation and multi-modal knowledge transfer for few-shot object detection. *Pattern Recognition*, *161*, Article 111283.

Du, S., Zhu, H., Xiong, G., et al. (2023). Semantic similarity information discrimination for video captioning. *Expert Systems with Applications*, *213*, Article 118985. http://dx.doi.org/10.1016/j.eswa.2022.118985, URL https://www.sciencedirect.com/science/article/pii/S0957417422020036.

Fang, Z., Gokhale, T., Banerjee, P., et al. (2020). Video2commonsense: Generating commonsense descriptions to enrich video captioning. arXiv preprint arXiv:2003.05162.

Fang, S., Wang, S., Zhuo, J., et al. (2022). Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 4789–4800).

Fei, M., Jiang, W., & Mao, W. (2021). Learning user interest with improved triplet deep ranking and web-image priors for topic-related video summarization. *Expert Systems with Applications*, *166*, Article 114036.

Feichtenhofer, C., Fan, H., Malik, J., et al. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* ICCV, (pp. 6202–6211).

Gautam, S., Sarkhoosh, M. H., Held, J., Midoglu, C., Cioppa, A., Giancola, S., et al. (2024). SoccerNet-Echoes: A Soccer game audio commentary dataset. arXiv preprint arXiv:2405.07354.

Gu, X., Chen, G., Wang, Y., et al. (2023). Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 18941–18951).

He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hou, J., Wu, X., Zhang, X., et al. (2020). Joint commonsense and relation reasoning for image and video captioning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10973–10980).

Huang, Y., Wang, Y., Zeng, Y., & Wang, L. (2022). MACK: multimodal aligned conceptual knowledge for unpaired image-text matching. *Advances in Neural Information Processing Systems*, *35*, 7892–7904.

Khan, H., Hussain, T., Ullah Khan, S., Ahmad Khan, Z., & Baik, S. W. (2024). Deep multi-scale pyramidal features network for supervised video summarization. *Expert Systems with Applications*, *237*, Article 121288.

Kim, B. J., & Choi, Y. S. (2020). Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th annual ACM symposium on applied computing* (pp. 1056–1065).

Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Li, P., Wang, T., Zhao, X., Xu, X., & Song, M. (2025). Pseudo-labeling with keyword refining for few-supervised video captioning. *Pattern Recognition*, *159*, Article 111176.

Li, Y.-L., Xu, L., Liu, X., et al. (2020). Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 382–391).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Lin, K., Li, L., Lin, C.-C., et al. (2022). Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 17949–17958).

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., et al. (2022). Video swin transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 3202–3211).

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., et al. (2020). Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353.

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

Ma, Y., Zhu, Z., Qi, Y., Beheshti, A., Li, Y., Qing, L., et al. (2024). Style-aware two-stage learning framework for video captioning. *Knowledge-Based Systems*, *301*, Article 112258.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Miech, A., Alayrac, J.-B., Smaira, L., et al. (2020). End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 9879–9889).

Mkhallati, H., Cioppa, A., Giancola, S., et al. (2023). SoccerNet-Caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 5073–5084).

Nabati, M., & Behrad, A. (2023). Multi-sentence video captioning using spatial saliency of video frames and content-oriented beam search algorithm. *Expert Systems with Applications*, *228*, Article 120454. http://dx.doi.org/10.1016/j.eswa.2023.120454.

Pan, B., Cai, H., Huang, D.-A., et al. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 10870–10879).

Papineni, K., Roukos, S., Ward, T., et al. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* ACL, (pp. 311–318).

Qi, M., Wang, Y., Li, A., et al. (2019). Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(8), 2617–2633.

Qi, J., Yu, J., Tu, T., et al. (2023). 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* (pp. 5391–5395).

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* ICML, (pp. 8748–8763). PMLR.

Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Rao, J., Wu, H., Jiang, H., Zhang, Y., & Xie, Y. W. W. (2024). Towards universal Soccer video understanding. arXiv preprint arXiv:2412.01820.

Rao, J., Wu, H., Liu, C., Wang, Y., & Xie, W. (2024). MatchTime: Towards automatic soccer game commentary generation. arXiv preprint arXiv:2406.18530.

Sap, M., Le Bras, R., Allaway, E., et al. (2019). Atomic: An atlas of machine commonsense for if-then reasoning. *vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3027–3035).

Shao, H., Fang, Z., & Yang, Y. (2022). CAVAN: Commonsense knowledge anchored video captioning. In *Proceedings of the 26th international conference on pattern recognition* ICPR, (pp. 4095–4102). IEEE.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2556–2565).

Shen, Y., Gu, X., Xu, K., Fan, H., Wen, L., & Zhang, L. (2023). Accurate and fast compressed video captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15558–15567).

Smith, E., & Kosslyn, S. (2007). *Cognitive psychology: Mind and brain*. Pearson/Prentice Hall.

Tang, M., Wang, Z., Liu, Z., et al. (2021). Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 4858–4862).

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 4566–4575).

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729.

Wang, J., Chen, D., Luo, C., He, B., Yuan, L., Wu, Z., et al. (2024). Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18209–18220).

Wang, L., Xiong, Y., Wang, Z., et al. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* ECCV, (pp. 20–36). Springer.

Wu, G., Song, S., Wang, X., & Zhang, J. (2024). Reconstructive network under contrastive graph rewards for video summarization. *Expert Systems with Applications*, *250*, Article 123860. http://dx.doi.org/10.1016/j.eswa.2024.123860, URL https://www.sciencedirect.com/science/article/pii/S0957417424007267.

Wu, D., Zhao, H., Bao, X., et al. (2022). Sports video analysis on large-scale data. In *European conference on computer vision* ECCV, (pp. 19–36). Springer.

Xi, Z., Shi, G., Li, X., Yan, J., Li, Z., Wu, L., et al. (2025). A simple yet effective knowledge guided method for entity-aware video captioning on a basketball benchmark. *Neurocomputing*, *619*, Article 129177.

Xiong, H., Wang, L., Qiu, H., Zhao, T., Qiu, B., & Li, H. (2025). Adaptively forget with cross-modal and textual distillation for class-incremental video captioning. *Neurocomputing*, Article 129388.

Xu, J., Huang, Y., Hou, J., et al. (2024). Retrieval-augmented egocentric video captioning. arXiv preprint arXiv:2401.00789.

Xu, W., Xu, Y., Miao, Z., Cen, Y., Wan, L., & Ma, X. (2025). CroCaps: A CLIP-assisted cross-domain video captioner. *Expert Systems with Applications*, *268*, Article 126296.

Xu, J., Yao, T., Zhang, Y., et al. (2017). Learning multimodal attention LSTM networks for video captioning. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 537–545).

Yang, B., Cao, M., & Zou, Y. (2023). Concept-aware video captioning: Describing videos with effective prior information. *IEEE Transactions on Image Processing*.

Yang, B., Zhang, T., & Zou, Y. (2022). Clip meets video captioning: Concept-aware representation learning does matter. In *Chinese conference on pattern recognition and computer vision* PRCV, (pp. 368–381). Springer.

Yao, L., Torabi, A., & Cho, K. (2015). Describing videos by exploiting temporal structure. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 4507–4515).

Ye, H., Li, G., Qi, Y., et al. (2022). Hierarchical modular network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 17939–17948).

Yu, H., Cheng, S., Ni, B., et al. (2018). Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6006–6015).

Yu, W., Liang, J., Ji, L., et al. (2021). Hybrid reasoning network for video-based commonsense captioning. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 5213–5221).

Yuan, M., Jia, G., & Bao, B.-K. (2023). GPT-based knowledge guiding network for commonsense video captioning. *IEEE Transactions on Multimedia*.

Zeng, Y., Wang, Y., Liao, D., et al. (2024). Contrastive topic-enhanced network for video captioning. *Expert Systems with Applications*, *237*, Article 121601. http://dx.doi.org/10.1016/j.eswa.2023.121601.

Zeng, P., Zhang, H., Gao, L., et al. (2023). Visual commonsense-aware representation network for video captioning. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, B., Gao, J., & Yuan, Y. (2024). A descriptive basketball highlight dataset for automatic commentary generation. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 10316–10325).

Zhang, Y., Liu, Y., & Wu, C. (2024). Attention-guided multi-granularity fusion model for video summarization. *Expert Systems with Applications*, *249*, Article 123568.

Zhang, Z., Qi, Z., Yuan, C., et al. (2021). Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 9837–9846).

Zhang, Z., Shi, Y., Yuan, C., et al. (2020). Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* CVPR, (pp. 13278–13288).

Zhou, H., Luo, T., & He, Y. (2025). Dynamic collaborative learning with heterogeneous knowledge transfer for long-tailed visual recognition. *Information Fusion*, *115*, Article 102734.

Zhuo, J., Zhu, Y., Cui, S., et al. (2022). Zero-shot video classification with appropriate web and task knowledge transfer. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 5761–5772).