STaR: Multi-Granular Spatio-Temporal Reasoning for Long-Form Dense Video Captioning

Anonymous Submission

Abstract. Dense video captioning is crucial for enhancing video understanding in daily applications and presents a significant challenge in multimodal analysis. Existing methods often overlook video-to-dynamic-space mapping at varying scales, resulting in captions that lack specificity and remain overly general, failing to capture real-world physical detail. To address this limitation, we propose a multi-granularity Spatio-Temporal Reasoning (STaR) approach, which integrates: (i) efficient global feature integration to model long-term temporal dependencies, (ii) spatial attention mechanisms with position encoding to capture absolute spatial information, and (iii) cross-modal feature fusion to align and unify global, local, and spatial representations. Moreover, we enhance the framework using a Large Language Model (LLM) to improve the richness and naturalness of the generated descriptions. Comparative experiments have been conducted to evaluate the effectiveness of the proposed method on SoccerNet dataset. Experimental results demonstrate that our model effectively enhances localization accuracy and generates captions with superior temporal and spatial detail fidelity. *The code will be available.*

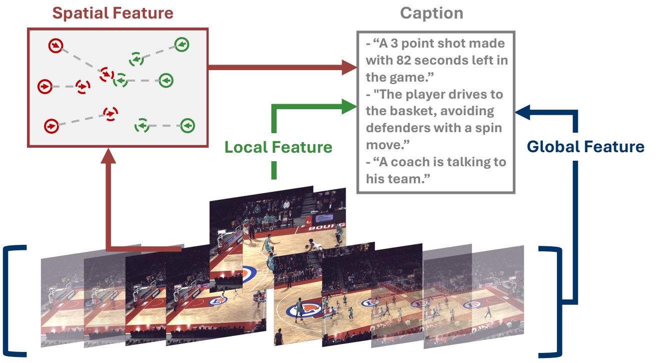

Spatial Feature
Caption
- "A 3 point shot made with 82 seconds left in the game."
- "The player drives to the basket, avoiding defenders with a spin move."
- "A coach is talking to his team."
Local Feature
Global Feature

Figure 1: **Enhanced Descriptions with Global Temporal and Spatial Context.** By incorporating global temporal context and absolute positional information into local segment visual features, our method produces more detailed and comprehensive descriptions.

1 Introduction

With the increasing focus on video understanding and multimodal analysis, video captioning has become a rapidly evolving research area. While conventional video captioning, which generates descriptions for pre-segmented clips, has shown promising results, it faces significant challenges when applied to dense video captioning. This task involves two key objectives: (1) localizing event segments in untrimmed videos and (2) generating natural language descriptions for these segments. Current methods are limited by insufficient fine-grained feature integration, which restricts their ability to capture detailed contextual information within events. Furthermore, modeling long-range temporal dependencies in extended videos remains a critical unsolved problem. Addressing these limitations—fine-grained feature fusion and long-term temporal modeling—is essential for advancing dense video captioning performance.

In recent years, various methods have been developed to extract fine-grained features from videos, enhancing performance in related tasks. For example, [7] integrates camera calibration with player localization to improve action recognition, while [31, 49] leverage the spatial distribution of players and objects for tactical analysis. Action recognition, as explored in [11], aims to accurately identify specific moments within untrimmed soccer videos. To address feature perception in long videos, several approaches have been proposed, including global-local representation modeling [37] to produce rich semantic video captions, context-aware loss functions [6] for capturing temporal relationships, spatiotemporal encoders [9] for leveraging temporal information, and Transformer-based [48] or anchor-based models [29] for processing long sequential data. Recent advancements also include leveraging external memory [17] to enhance long-term feature representation and scalable caption producers [4] to generate high-quality captions.

However, existing video captioning methods still face significant challenges in modeling hour-level temporal inputs and lack effective mechanisms to efficiently align fine-grained and global features: 1) **Macro-level Temporal Context.** Current methods in Dense Video Captioning (DVC) predominantly rely on localized, context-restricted visual features for temporal localization and caption generation. While these approaches are effective for short-duration videos (*i.e.*, ≤ 3 minute) with limited captioning demands, they exhibit significant inefficiencies when applied to extended video sequences, such as those spanning several hours. The failure to capture the broader temporal context within which local segments are embedded results in captions that lack a coherent sense of global temporal structure. Furthermore, for long videos, existing methods [14] typically generate captions by extracting features from short and sparsely sampled segments, which are then aggregated onto longer segments (using low-level captions to enhance high-level descriptions). In contrast, our method simultaneously perceives both local and global visual features, enabling a more comprehensive understanding of the temporal context. 2) **Spatial Feature Deterioration.**

Traditional caption generation methods often suffer from spatial feature degradation, where key spatial information (e.g., location data of key entities) is sparse and diluted by less relevant scene details. Limited inter-frame visual variations further complicate accurate spatial feature extraction, resulting in loss of contextual nuances in generated captions. To address this issue, our framework introduces a fine-grained feature enhancement integration approach. In the caption generation stage, it integrates local features, absolute position features, and global features, and aligns them with a large language model (LLM) input to obtain a coherent output.

Our method specifically addresses the challenges of macro-level temporal context perception and spatial feature degradation by a Multi-Granular Spatio-Temporal Reasoning (STaR) method through the following approaches: **First**, a latent learning framework integrates variable-length global features with fixed-length local features into a unified representation, enabling the capture of broader temporal context. This dual-scale framework ensures that the captions can capture visual features beyond the limitations of local windows while maintaining broader temporal consistency. **Second**, a Spatial Semantic Synthesizer (SSS) module extracts positional data and encodes it through multi-resolution hash encoding. These fine-grained features are fused with local features and further integrated with global features to preserve critical spatial information. **Third**, The Global Context Capturing (GCC) and Position Context Capturing (PCC) modules are employed to dynamically refine multi-modal features, specifically tailored for the spotting and caption generation tasks, respectively. These modules effectively condense the enhanced information into a set of tokens that can be integrated and processed by the Large Language Model (LLM). Our framework enhances caption quality by mitigating the degradation of spatial features and ensuring the retention of fine-grained details as well as broader contextual coherence.

Specifically, our key contributions are threefold:

- We propose STaR, a novel DVC method that preserves fine-grained spatio-temporal features by leveraging dynamic position encoding. This is the first approach to integrate view-invariant spatial features (which unify directional information across camera perspectives) with temporal dynamics, enabling robust modeling of real-world scene variations in DVC tasks.
- We propose a captor architecture with multi-granular attention mechanisms, termed Context Capturer (CC), enabling the efficient combination of global, local, and spatial semantic video features for specific events. This design ensures adaptability to varying video durations within a fixed event window.
- Our method offers two primary advantages: First, it improves interpretability for long-form videos ($\geq$ 45 min), enhancing generalizability to daily videos. Second, it accommodates the intersectional reasoning of flexible temporal events, enabling denser caption generation.

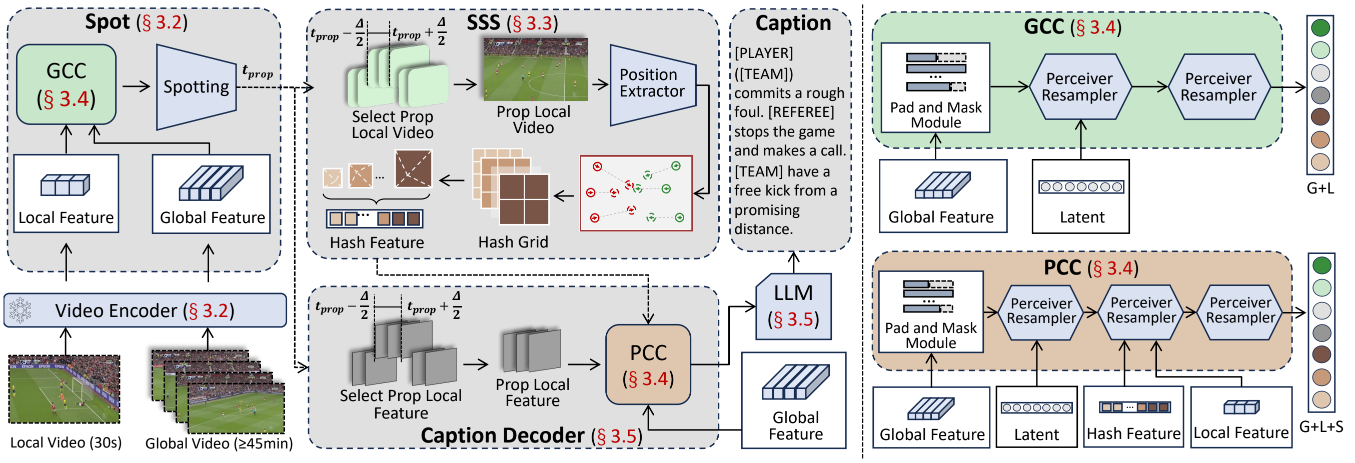
STaR represents a highly intuitive yet versatile captioning framework, it is compatible with different video-language network architectures. Specifically, it surpasses the previous state-of-the-art DVC method, SDVC [21], by ↑ **13.84** $\sim$ **16.81** in CIDEr, ↑ **1.23** $\sim$ **3.31** in METEOR, and ↑ **1.43** $\sim$ **3.44** in BLEU4. In **§4.3**, STaR further compares the performance of caption generation models by using ground truth timestamps to eliminate the influence of localization models, achieving superior results across all metrics (at least ↑ **1.49** higher in B@4, ↑ **1.57** higher in METEOR, ↑ **5.45** higher in CIDEr, and ↑ **2.35** higher in ROUGE-L). Importantly, this performance is achieved without the necessity of pre-training on large-scale, related datasets, underscoring its generalization potential. We hope this work could bring fundamental insights into related fields. Our code will be released.

2 Related Work

Video Captioning. Video captioning aims to generate textual descriptions for videos, bridging the gap between visual content and natural language. Previous advancements mainly focus on unified frameworks, such as pre-trained models like Flamingo [1] and Frozen [2], which integrate video and language modalities for end-to-end optimization. One-stage approaches, inspired by YOLO [27], including [33] and Vid2Seq [39], unify localization and captioning tasks to improve efficiency. Recently, more advanced architectures focus on innovative modules, such as global-local representation [36], stacked multimodal attention [44], two-step transformer-based polishing network [35], external memory [17], and scalable captions producer [4]. Additionally, [42] achieve significant advancements by precise spatial computations. Despite these advancements, most methods still struggle to capture complex event relationships and temporal dependencies in hour-level videos. However, they often lack mechanisms to maintain narrative consistency over extended temporal spans, motivating our global-local feature fusion approach.

Dense Video Captioning. Dense video captioning (DVC) addresses the dual task of temporal event localization and corresponding caption generation. The first framework by [18] employed a multi-scale proposal module and an attention-based LSTM, establishing the "detect-then-describe" paradigm. Subsequent improvements introduced context modeling [32, 40], reinforcement learning [20, 22], and multi-modal feature fusion [13, 12]. Recent advancements in DVC include event proposal generation, feature extraction, and aggregation [43, 34], attention mechanisms [25], end-to-end optimization [2], vision–language alignment [5], sub-global feature fusion [10], prompt tuning [38], and weak supervision [47]. However, these methods rely on local temporal windows and often miss global context, whereas our BLIP3-inspired [19] structure efficiently captures information across the entire video. Overall, current DVC approaches combine multi-scale proposals, attention-based context modeling, reinforcement learning, and multimodal fusion to improve temporal localization and caption quality, yet they continue to face challenges in comprehensive video-level understanding.

Fine-grained Feature Enhancement. Fine-Grained features refer to detailed classification or description of objects, often utilized in video captioning to depict actions and interactions. [7] enhances performance by integrating camera calibration with player localization. [41] proposed a three-branch network for spatio-temporal localization and relationship modeling, whereas [28] utilizes coarse- and fine-grained matching with pre-trained embeddings. Additionally, [31, 49] leveraged player and object spatial positioning for tactical analysis, while [24] proposed a multi-hot vector-based length embedding technique. However, these methods primarily extract sparse spatial features, limiting their effectiveness in videos with significant viewpoint changes. To address this, we calibrate the camera for absolute positions and use Hash Grid to densify features, facilitating gradient descent optimization. Our proposed camera calibration and Hash Grid densification further ensure robust, view-invariant feature representation, addressing the sparsity issues inherent in previous fine-grained methods.



(a) Overall Architecture

(b) GCC and PCC modules

Figure 2: **The framework of our proposed method.** Fig. 2(a) illustrates the overall architecture. A frozen anchor-associated video encoder pre-extracts spatio-temporal features from the video, while a spot module classifies local features to determine caption positions and categories. The GCC module enhances localization by integrating local and global information. The SSS module utilizes hash grid encoding to extract positions and learn sparse location features. Local video selection is focused on a Δ-sized window around the previously localized position t_{prop}. The PCC module fuses spatial, global, and local features, aligns them with a large language model (LLM), and generates captions. As shown in Fig. 2(b): 1) The **GCC** module receives local features, global features, and latents, where latents, as learnable vectors, act as query vectors in the Perceiver Resampler to compress and extract global feature information, producing latents containing both global and local features (G+L). 2) The **PCC** module uses local features, global features, spatial features, and latents as inputs to extract spatial information via the Perceiver Resampler, resulting in latents enriched with spatial, global, and local features (G+L+S).

3 Method

Our goal is to improve Dense Video Captioning (DVC) by integrating fine-grained spatial features and long-term temporal context, enabling the model to generate more precise descriptions. For this, we propose a Multi-Granular Spatio-Temporal Reasoning (STaR) framework integrates multi-modal features to produce detailed and contextually accurate video descriptions. In our framework, the Context Capturer (CC) integrates global and local video features, aligning them with the inputs of the Large Language Model (LLM). In parallel, the Spatial Semantic Synthesizer (SSS) enhances video features by encoding positional information through a hashing encoding, generating hash-encoded features for supplementary semantic enrichment. Our model architecture, illustrated in Fig. 2, follows a two-stage framework inspired by the SDVC baseline [21], which consists of a spotting phase and a captioning phase.

3.1 *Preliminaries*

The *Perceiver Resampler* [19] is an encoder architecture based on perceiver attention, designed to process high-dimensional visual features of variable lengths and align multimodal data. Given visual features $X_f \in \mathbb{R}^{t\times s\times d}$ and time embeddings $TE \in \mathbb{R}^{t\times 1\times d}$, where t, s, and d denote temporal, spatial, and feature dimensions, the visual features are flattened and combined with time embeddings. These are concatenated with learned latents $Z \in \mathbb{R}^{r\times d}$ and projected to obtain query (q), key (k), and value (v) vectors:

$$\tilde{X} = \text{Concat}(X_f, Z) \in \mathbb{R}^{(ts+r)\times d}, \tag{1}$$

where $X_f \in \mathbb{R}^{ts\times d}$ represents flattened visual features, and $Z \in \mathbb{R}^{r\times d}$ denotes learned latents. The linear map are defined as:

$$\begin{aligned} q &= (ZW_q), \\ k &= (\tilde{X}W_{kv})[:, :(nd)_{\text{head}}], \\ v &= (\tilde{X}W_{kv})[:, (nd)_{\text{head}}:], \end{aligned} \tag{2}$$

where $W_q \in \mathbb{R}^{d\times(nd)_{\text{head}}}$ and $W_{kv} \in \mathbb{R}^{d\times 2(nd)_{\text{head}}}$ are learnable projection matrices. The perceiver attention output is computed via cross-attention:

$$X'_f = \text{CrossAttn}(q, k, v). \tag{3}$$

Here, k and v are derived from $\tilde{X}$, while q is obtained from the learned latents Z. Summarize the above three operations as the perceiver attention mechanism and write it as a formula:

$$Z' = \text{PerceiverAttn}(X_f, Z). \tag{4}$$

The Perceiver Resampler stacks multiple perceiver attention layers, each followed by a Feed Forward (FF) layer, functioning as a decoder. This is formally expressed as:

$$Z' = \text{PerceiverResampler}(X_f, Z), \tag{5}$$

where Z' represents a fixed number of output tokens encoding X_f in compressed form. A variant of this module includes self-attention. The subsequent *Context Capturer* builds on these two architectural variants.

3.2 *Anchor-Associated Video Encoder*

According to the task definition of *Single-anchored Dense Video Captioning* (SDVC) [21], given a long, untrimmed video V, our goal is to generate a temporally anchored caption C_i at each timestamp t_i, where $i \in \{1, 2, \ldots, N\}$ and N is the total number of key moments in the video. Formally, the task can be expressed as:

$$C_i = f(V, t_i) \tag{6}$$

where $f(\cdot)$ is the caption generation model that takes the video V and a specific timestamp t_i, and outputs a natural language description C_i of the event occurring around that timestamp. The SDVC task is sparser than traditional DVC [18], which generates captions for temporally bounded intervals; instead, SDVC focuses on generating a single caption for each key moment in the video and uses a single timestamp for spotting.

However, the caption associated with each anchor is not solely relevant to its corresponding fragment but also exhibits contextual relationships with adjacent fragments and, in some cases, the entire video. Thus, to obtain a more comprehensive representation at each timestamp, we employ a pre-trained video encoder to extract spatio-temporal features X_v over varying temporal range within each video. This anchor-associated video encoder excels at capturing spatial and temporal dynamics and modeling inter-frame dependencies within a sequence. When processing local videos and global videos independently, distinct temporal correlations are derived. This suggests that removing local features from global video features before feeding them into subsequent modules would disrupt the inherent frame dependencies in the global features. To boost training efficiency and minimize I/O overhead, the video encoder is frozen during training.

3.3 Spatial Semantic Synthesizer (SSS)

In this section, we introduce the *Spatial Semantic Synthesizer*, which enhances the visual features of videos by incorporating spatial information about the subjects in the scene. This module leverages existing object localization techniques [15] to extract the positions of visible subjects and encodes them using a hash encoding scheme [23]. The resulting features, which vary in length depending on the number of subjects in the frame, are then used to augment the visual features of the current frame through an attention mechanism.

Given a video frame, we detect and localize all visible subjects using a pre-trained object detection model, representing each subject's position as a coordinate $(x_1, x_2, \ldots, x_n)$. To encode these sparse positions into a feature space, we apply a hash encoding function, mapping each subject's spatial location to a fixed-dimensional feature vector $p_i \in \mathbb{R}^d$, where d is the feature dimensionality. The set of all encoded subject positions in the frame is denoted as $P = \{p_i\}_{i=1}^{N}$ where N is the number of subjects visible in the frame.

To integrate subject features with frame visual features, we use a perceiver attention mechanism. For denser feature aggregation and computational efficiency, attention is applied in a low-dimensional space ($d' < d$), with linear transformations W_{down} and W_{up} for dimension alignment. The enhanced visual feature $\tilde{X}'_v$ is then computed as PerceiverAttn$(P, X_v W_{down})$ where the visual feature X_v is the local video feature selected in the spotting stage and the spatial features P serve as keys and values while $X_v W_{down} = X'_v \in \mathbb{R}^{d'}$ serve as queries for the attention input.

Finally, the projection is mapped back to the high-dimensional space, ensuring alignment with the original dimensionality of the extracted image features X_v. This augmentation is represented as: $\tilde{X}_v = \tilde{X}'_v W_{up} + X_v$. The spatial feature extraction module can thus be formally expressed as:

$$H = \text{SSS}(X_v) \tag{7}$$

where H is the predicted hash-encoding absolute locations of subjects in the real-world space.

The SSS effectively integrates spatial information about subjects into frame-level visual features through an attention-based mechanism. This enables the caption generation module to learn the overall position features from the sparse and detailed feature representation.

3.4 Context Capturer (CC)

We introduce the *Context Capturer* module, which extracts and integrates global and local video features around a single anchor point, aligning them with spatial features. The module consists of two components: the *Global Context Capturer (GCC)*, which performs perceptive resampling of varying-length global video features to match the length of local features before fusion; and the *Position Context Capturer (PCC)*, which aligns hash position features with local features for modal alignment [19].

Global Context Capturer. This module addresses the challenge of fusing global video features of varying lengths with local video features of fixed length. Specifically, given global video features G, where each feature has varying lengths, we pad them to a fixed length and generate masks to distinguish meaningful elements from padding. The global video feature set is expressed as $G = \{g_i\}_{i=1}^{n}$, where n represents the number of videos. For each video, standard caption timestamps are defined in the dataset as $T_j = \{t_i^{cap}\}_{i=1}^{N}$.

We define a function h that extracts timestamps from the global video feature g_i and maps them to the set T_j as $h : g_i \rightarrow T_j$. The global video feature g_i is then trimmed into a series of clips around each caption anchor t_n^{cap}, represented as local video features:

$$L = h(g_i) = \{l_i\}_{i=1}^{m} \tag{8}$$

where L is the set of local features extracted from the global feature g_i to generate captions.

To extract and fuse global and local video features, we use learnable latents $Z \in \mathbb{R}^{n_{heads} \times d}$, where n_{heads} is the number of compressed tokens, which constitute the learnable latents. Initially, the latents Z and padded global features G are input into the Perceiver Resampler to incorporate global information. Then, Z and local features L are fed into the Resampler to encode local information. The process is summarized as:

$$Z' = \text{GCC}(G, L, Z) \tag{9}$$

where Z' denotes the output tokens. See Fig. 2 for details. At this stage, Z have successfully integrated both local and global information. Our method can effectively capture and combine both global and local perspectives of the video content, thus providing richer understanding for both the localization subtask and the caption generation subtask in the overall task.

Position Context Capturer. This module addresses the challenge of aligning local position features with the hash video features extracted from the SSS module. A similar approach to the fusion of global and local information is used, employing a learnable latent to align relevant parts of the hash position features with the local video features. Specifically, given $SSS(X_v) = H \in \mathbb{R}^{t \times a \times d_{hash}}$, H represents the hash position features, where a is the maximum number of people in the frames, d_{hash} is the dimension of the hash encoding output, and X_v is the original video.

To align the hash position feature H with the local video features L, we use a structured approach represented as $L' =$

Method	Encoder	F	B4@30	M@30	R-L@30	C@30	Recall	Precision	F1
Vid2Seq [39]	CLIP	1	6.69	11.31	8.82	26.33	20.05	49.15	28.48
PDVC [33]	C3D	2	0.79	0.83	0.61	2.13	1.23	1.12	1.17
SDVC [21]			**24.77**	20.93	33.47	19.02	55.66	**76.15**	**64.31**
Delta [8]			23.65	22.48	33.80	29.29	46.95	74.43	57.58
Ours (STaR)			23.98	**24.29**	**34.28**	**33.60**	**55.80**	74.60	63.84
PDVC [33]	I3D	2	0.83	0.88	0.64	1.86	1.53	1.24	1.37
SDVC [21]			21.98	18.52	28.79	8.65	53.61	73.41	61.97
Delta [8]			23.82	20.79	31.89	19.43	37.23	73.15	49.35
Ours (STaR)			**23.96**	**24.17**	**33.34**	**30.52**	**55.97**	**74.70**	**63.99**
PDVC [33]	ResNET	2	0.88	1.13	0.64	1.75	1.97	1.35	1.60
SDVC [21]			24.70	21.31	33.45	18.30	**60.31**	75.85	**67.19**
Delta [8]			**24.86**	22.62	34.34	29.46	42.50	**76.38**	54.61
Ours (STaR)			24.73	**24.98**	**34.75**	**33.72**	56.69	74.64	64.44
PDVC [33]	Baidu	1	1.12	1.50	0.81	2.48	1.92	1.63	1.76
SDVC [21]			21.17	21.99	27.01	27.63	35.21	84.32	49.68
Delta [8]			23.18	24.07	**35.73**	24.66	13.94	**85.53**	23.97
Ours (STaR)			**24.61**	**25.30**	35.32	**41.47**	**55.94**	75.42	**64.24**

Table 1: **Comparisons with state-of-the-art methods** on SoccerNet-Caption dataset. F denotes the frame rate of the extracted video features in the original video. B4/M/R-L/C is short for BLEU4/METEOR/ROUGE-L/CIDEr. @30 means the window size used to generate caption. Baidu encoder denotes the TimesFormer model trained by Baidu.

Dataset	Video	Avg Duration	Avg Sentences
ActivityNet	20k	180 (s)	5
YouCook2	2k	30 (s)	7.7
SoccerNet	942	2735.9 (s)	39.2

Table 2: **Comparison of SoccerNet-Caption with other captioning datasets.** The SoccerNet dataset features the longest individual video duration and the highest annotation density within a single video, demonstrating that SoccerNet-Caption is a great benchmark for research on dense video captioning.

PerceiverResampler(H, L), where L' represents the most relevant part of H to L, extracted via the cross-attention mechanism in the Perceiver Resampler.

To mitigate dataset imbalance—which causes salient visual cues to be highly sparse (as shown by the ablation results in Table 3)—we first apply a *Filter* module parameterized by a ratio α that uniformly samples a subset of frames. Formally, given the input feature sequence $G = \{g_1, \ldots, g_T\}$, we partition it into two disjoint subsets:

$$G_\alpha,\, G_{1-\alpha} = \mathrm{Filter}_\alpha(G), \tag{10}$$

where $\alpha \in (0, 1)$ is the sampling ratio (sampling interval $k = 1/\alpha$, typically $\alpha = 0.2$).

To recover potentially informative frames omitted by this coarse filter—especially in long or rare events—we introduce a *Memory* module. We compute cosine similarity between each frame in $G_{1-\alpha}$ and the retained set G_α; frames whose maximum similarity falls below a predefined threshold τ are deemed informative and re-inserted into the filtered set. The enriched feature set becomes

$$G' = G_\alpha \,\cup\, \{\, g \in G_{1-\alpha} \mid \max_{g' \in G_\alpha} \cos(g, g') < \tau \}. \tag{11}$$

Finally, G' is fed into the Position-wise Cross-modal Calibration (PCC) module together with local context L' and additional features Z, yielding the aligned output

$$Z' = \mathrm{PCC}(G',\, L',\, Z). \tag{12}$$

See Fig. 2 for an overview of the entire pipeline.

3.5 *Large-Language-Model Caption Decoder*

In this work, we adopt a trainable large language model (LLM) as the caption decoder, enabling seamless fusion of video understanding and language generation. To integrate multimodal inputs, we inject two lightweight adapter modules into each Transformer block of the LLM: the Spatial Semantic Synthesizer synthesizes spatial features within each frame, reinforcing relationships between objects and finer-grained visual semantics; the Context Capturer aligns motion dynamics across time with the decoder's textual context, capturing macro-level temporal information.

By integrating detailed spatial information from SSS and temporal context from CC, the decoder generates precise captions describing object positions and action sequences. The entire framework—including input projections, adapters, and the base LLM—is end-to-end trainable using token-level cross entropy loss. During inference, captions are generated autoregressively via beam search.

4 Experiments

4.1 *Experimental Settings*

Implementation Details. Our video encoder is based on Baidu's PP-TimesFormer [46, 3], pre-trained on K400 [16] and fine-tuned on SoccerNet [11]. To reduce memory and computational costs, we use a frozen model and lower the video frame rate to 1 FPS before inference. The position extractor comes from the sn-gamestate project [30], using its pre-trained model, with player position data pre-computed for efficiency. The LLM decoder is built on GPT-2 medium [26], which remains trainable during training. In the overarching model, visual features are projected into a 768-dimensional latent space. Two Context Perceiver Resampler layers integrate global and spatial features. The hash position encoder within the SSS module employs a HashGrid encoder coupled with a FullyFusedMLP, generating 24-dimensional feature representations for a maximum of

Blurring	SSS	B4@30	M@30	R@30	C@30	B@4	M	R	C
30	×	24.74	20.80	32.83	23.78	4.12	22.19	21.54	18.50
30	✓	24.13$\downarrow_{0.61}$	24.90$\uparrow_{\mathbf{4.10}}$	34.69$\uparrow_{\mathbf{1.86}}$	35.43$\uparrow_{\mathbf{11.65}}$	7.07$\uparrow_{\mathbf{2.95}}$	24.66$\uparrow_{\mathbf{2.47}}$	24.87$\uparrow_{\mathbf{3.33}}$	32.25$\uparrow_{\mathbf{13.75}}$
10	×	24.14	23.64	34.02	29.68	6.74	23.97	25.07	29.12
10	✓	24.44$\uparrow_{0.30}$	24.81$\uparrow_{1.17}$	34.97$\uparrow_{0.95}$	36.98$\uparrow_{7.30}$	7.18$\uparrow_{0.44}$	24.28$\uparrow_{0.31}$	25.00$\downarrow_{0.07}$	30.65$\uparrow_{1.53}$
3	×	22.86	24.42	34.18	32.36	7.02	24.10	24.56	26.99
3	✓	24.16$\uparrow_{1.30}$	24.96$\uparrow_{0.54}$	35.02$\uparrow_{0.84}$	35.03$\uparrow_{2.67}$	7.27$\uparrow_{0.25}$	24.51$\uparrow_{0.41}$	25.44$\uparrow_{0.88}$	30.16$\uparrow_{3.17}$
1	×	22.21	24.44	34.17	35.22	6.86	24.39	25.52	31.03
1	✓	24.61$\uparrow_{\mathbf{2.40}}$	25.30$\uparrow_{0.86}$	35.32$\uparrow_{1.15}$	41.47$\uparrow_{6.25}$	7.63$\uparrow_{0.77}$	24.68$\uparrow_{0.29}$	25.72$\uparrow_{0.20}$	33.85$\uparrow_{2.82}$

Table 3: **Ablation results of artificial spatial feature deterioration.** Blurring denotes the feature homogenization window size. The performance is measured by degree of artificial spatial feature degradation of visual features. Feature homogenization is achieved by dividing the visual features according to the Blurring window size and selecting the first frame to cover the other features in the window.

Method (GT)	B4	M	R-L	C
Vid2Seq	6.93	10.58	8.67	24.02
SDVC	6.04	23.46	23.32	17.56
Delta	6.14	23.11	23.37	28.40
Ours	**7.63**	**24.68**	**25.72**	**33.85**

Table 4: **Performance of the captioning subtask in SoccerNet-Caption dataset.** GT denotes that the localization component of the input to the caption generation model utilizes ground truth proposals.

Method	ActivityNet †			YouCook2 †		
	B4	M	C	B4	M	C
PDVC	1.96	8.08	28.59	0.80	4.56	22.7
Ours (STaR)	**8.32**	**11.17**	**30.97**	**1.47**	**5.08**	**30.49**

Table 5: **Performance of Dense Video Captioning in ActivityNet Captions and YouCook2.** † denotes that the dataset is a subset extracted using similarity-based criteria.

G.	P.	B4@30	M@30	R-L@30	C@30	F1
×	×	23.61	23.81	34.41	35.88	64.88
✓	×	22.21	24.44	34.11	35.22	61.56
×	✓	24.29	24.09	34.50	34.41	**67.33**
✓	✓	**24.61**	**25.30**	**35.32**	**41.47**	64.24

Table 6: **Ablation results of different components.** G. denotes that the model uses global features, and P. denotes that the model uses positional features.

32 players per frame, thereby producing an output of dimensionality 24 × 32. The dimension of the latent is set to d = 768. The number of the latent tokens is set to 8. The max length of the text sequence is set to 100. For specific Settings, please refer to the table 1.

Dataset. We evaluate the proposed STaR using SoccerNet-Caption [11], a large-scale commentary benchmark dataset. It comprises 942 long uncut soccer match videos, averaging 2735 seconds per video, with 39.2 temporally localized sentences per video. Each sentence includes a context-dependent label for rough classification. Following the standard split, we use 562/184/196 videos for training, validation, and testing, respectively. In line with SoccerNet's format, we employ de-identified captions as generation targets, where specific titles are replaced with non-specific placeholders. For more details of the dataset comparison, please refer to Table 2.

To further expand the evaluation scope, we sample approximately 10% of the data from ActivityNet [18] and YouCook2 [45], selecting subsets that exhibit analogous characteristics to SoccerNet (such as prolonged video durations, high event density, and extended logical event chains) for cross-dataset testing.

4.2 Main Results

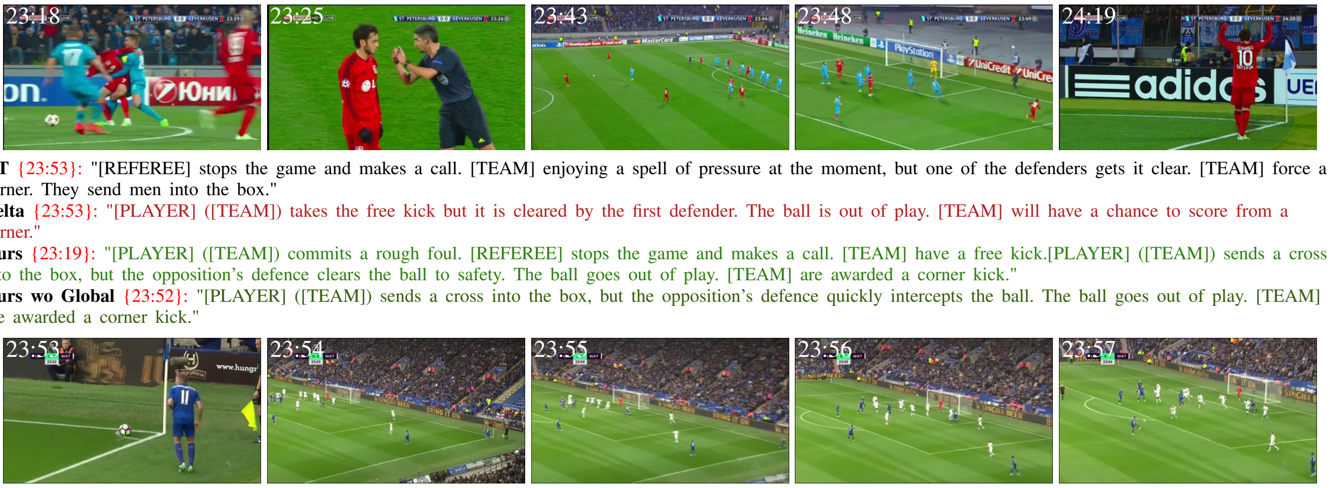
Comparison with State-of-the-Arts. Initially, we evaluated our approach on the standard benchmark dataset, SoccerNet Caption, employing the complete localization captioning pipeline. Table 1 presents a comparison of various methods across BLEU_4, METEOR, ROUGE-L, CIDEr, recall, precision, and F1 metrics, demonstrating the superior performance of our method across all indicators. Among the best performing Baidu encoder, the results indicate the following: a) Compared to the Delta method, which relies solely on local video features without the integration of long-range and spatial information, our method outperforms Delta with an increase of 1.43 in B4@30, 1.23 in M@30, and a significant 16.81 in C@30. b) Notably, our model excels particularly in C@30, suggesting that the output of our model is more natural and closer to human-like descriptions. c) Our model also performs well on visual features obtained by other encoders. As shown in Fig. 3, the prediction of the Delta model is often affected by the visual features beyond the window and the insufficient representation of spatial features, resulting in incorrect captions. STaR can overcome these challenges and obtain a comprehensive and accurate description. (d) The suboptimal performance of PDVC and Vid2Seq likely stems from two key factors. First, their training videos are significantly shorter than those in the DVC task, leading to a mismatch between their model parameters/architectures and the demands of long-form video processing. Second, the extended caption length in this dataset requires a more sophisticated decoder architecture to generate coherent and detailed descriptions effectively.

Furthermore, in Table 5, we extend our evaluation to the ActivityNet Caption and YouCook2 datasets. The experimental results demonstrate that our method achieves superior performance compared to previous dense video captioning approaches, achieving competitive results in cross-domain scenarios.

4.3 Ablation Study

We conduct ablation experiments on the SoccerNet caption benchmark to assess the capability of the proposed model.

Macro-Level Temporal Context Challenge. For complex movements beyond 30 seconds, STaR leverages global visual information from Local Video Feature Windows, while Delta struggles to capture



GT {23:53}: "[REFEREE] stops the game and makes a call. [TEAM] enjoying a spell of pressure at the moment, but one of the defenders gets it clear. [TEAM] force a corner. They send men into the box."

Delta {23:53}: "[PLAYER] ([TEAM]) takes the free kick but it is cleared by the first defender. The ball is out of play. [TEAM] will have a chance to score from a corner."

Ours {23:19}: "[PLAYER] ([TEAM]) commits a rough foul. [REFEREE] stops the game and makes a call. [TEAM] have a free kick.[PLAYER] ([TEAM]) sends a cross into the box, but the opposition's defence clears the ball to safety. The ball goes out of play. [TEAM] are awarded a corner kick."

Ours wo Global {23:52}: "[PLAYER] ([TEAM]) sends a cross into the box, but the opposition's defence quickly intercepts the ball. The ball goes out of play. [TEAM] are awarded a corner kick."

GT {23:52}: "[PLAYER] ([TEAM]) produces a lovely ball into the penalty area but the defender manages to intercept and comfortably averts the danger. The referee signals a corner kick to [TEAM]."

Delta {23:38}: "[PLAYER] ([TEAM]) takes the corner but fails to find any of his teammates. The ball is easily cleared out of danger."

Ours {23:36}: "[PLAYER] ([TEAM]) latches on to a precise defence-splitting pass and fires a nice strike from a promising distance towards the bottom left corner. [PLAYER] has a clear sight of it and makes a comfortable save. The ball is out of play and the linesman points at the corner flag. Corner to [TEAM]."

Ours wo Position {23:36}: "[PLAYER] ([TEAM]) receives a precise pass and is given time inside the box to send a promising strike towards the middle of the goal. [PLAYER] guesses his intention and denies him with a fine save."

Figure 3: **Detailed qualitative analysis of caption generation results.** In the first example, our model accurately describes the sequence leading to a corner kick, including the foul, referee's decision, free kick, cross, defensive clearance, and corner, aligning with the ground truth. The Delta model mentions a free kick resulting in a corner but omits the foul and defensive actions, while the variant without global features lacks context on the foul and free kick, producing a less informative description. In the second example, our model leverages spatial features to provide precise descriptions of player positions and actions, closely matching the ground truth, such as a pass attempt from outside the box and the ensuing corner kick. In contrast, the model without spatial features generates less detailed and occasionally inaccurate descriptions, underscoring the importance of spatial features for coherent and accurate commentary.

game progression. Qualitative analysis in Fig. 3 validates the effectiveness of hour-level context for long temporal spans. Compared to scenarios without global visual information, STaR generates more accurate captions, overcoming local window limitations.

Spatial Feature Deterioration Challenge. STaR uses Positional Prompts to produce precise captions reflecting video semantics. Delta fails to respond to subtle positional changes, resulting in incoherent captions. To mitigate Spatial Feature Deterioration, we demonstrate the role of Spatial Semantic Supplementation Synthesizer in alleviating dynamic feature collapse. Removing SSS significantly degrades performance, as shown in Table 3. We introduce blurring size K, where K consecutive frames are replaced with the first frame to simulate feature collapse. As K increases, SSS's impact on CIDEr and METEOR metrics becomes more pronounced. Evaluation in Table 3 shows SSS enables accurate captioning of detailed positions, even under severe feature collapse (large K) and limited visual information.

Spotting Subtask. Table 1 illustrates the quantitative evaluation for the localization subtask. STaR demonstrates remarkable robustness and balance in performing localization tasks across features from different encoders. Specifically, our method avoids extreme imbalances in the key metrics of Recall and Precision. Extremely low Recall would result in an insufficient number of clips being identified for caption generation, leading to incomplete and distorted evaluation of subsequent captioning metrics. On the other hand, extremely low Precision would cause the identified video clips to contain excessive redundant information, generating captions with little substantive content. In contrast, STaR effectively balances both the comprehensiveness and accuracy of spotting, thereby providing robust support for the subsequent caption generation task.

Captioning Subtask. Table 4 shows the quantitative evaluation for the caption generation subtask, with ground truth caption anchors provided. STaR outperforms others across all captioning metrics, showcasing its ability to accurately locate and comprehend video segments while generating precise descriptions. The substantial lead in CIDEr scores further emphasizes the naturalness of our model's language generation.

Spatio-Temporal Reasoning. As demonstrated in Table 6, STaR integrates global and positional features, outperforming other models in key metrics, especially F1 and CIDEr scores. This highlights the importance of these features in accurately identifying significant video events, improving the model's ability to understand and describe video content, and producing more natural and precise captions. Additionally, it emphasizes the synergistic effect of these features in boosting the model's overall performance.

5 Conclusion

This work presents a detail-oriented framework for long-form dense video captioning, addressing feature degradation and long-range perception issues through weighted fusion of low and high informative features. By enhancing semantically critical elements while balancing granular details and contextual information via spatial-temporal feature integration, the method achieves SOTA performance in event localization and caption generation. The framework's extensibility allows direct incorporation of domain-specific fine-grained features in subsequent work, though computational complexity in feature extraction remains a key challenge to address. Building on these results, Future work will explore extensions to streaming dense video captioning for real-time applications.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022. URL http://arxiv.org/abs/2204.14198.

[2] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1708–1718, Montreal, QC, Canada, 2021. IEEE.

[3] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, July 2021.

[4] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, L. Yuan, Y. Qiao, D. Lin, F. Zhao, and J. Wang. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. *arXiv preprint arXiv:2406.04325*, 2024.

[5] W. Chen, J. Niu, and X. Liu. MRCap: Multi-modal and Multi-level Relationship-based Dense Video Captioning. In *ICME*, 2023.

[6] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck, R. Gade, and T. B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *CVPR*, pages 13123–13133, Seattle, WA, USA, June 2020.

[7] A. Cioppa, A. Deliege, S. Giancola, F. Magera, O. Barnich, B. Ghanem, and M. V. Droogenbroeck. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *CVPRW, CVsports*, pages 4532–4541, Nashville, TN, USA, June 2021.

[8] A. Cioppa, S. Giancola, V. Somers, V. Joos, F. Magera, J. Held, S. A. Ghasemzadeh, X. Zhou, K. Seweryn, M. Kowalczyk, et al. Soccernet 2024 challenges results, 2024. URL https://arxiv.org/abs/2409.10587.

[9] A. Darwish and T. El-Shabrway. STE: Spatiotemporal encoder for action spotting in soccer videos. In *MMSports*, pages 87–92, Lisbon, Portugal, October 2022. ACM.

[10] Q. M. Dinh, M. K. Ho, A. Q. Dang, and H. P. Tran. Trafficvlm: A controllable visual language model for traffic video captioning. In *CVPR*, pages 7134–7143, 2024.

[11] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *CVPRW*, June 2018.

[12] V. Iashin and E. Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *BMVC*, 2020.

[13] V. Iashin and E. Rahtu. Multi-modal dense video captioning. In *CVPRW*, pages 4117–4126, 2020.

[14] M. M. Islam, N. Ho, X. Yang, T. Nagarajan, L. Torresani, and G. Bertasius. Video recap: Recursive captioning of hour-long videos. *arXiv preprint arXiv:2402.13250*, 2024.

[15] V. Joos, V. Somers, and B. Standaert. TrackLab. https://github.com/TrackingLaboratory/tracklab, 2024.

[16] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. URL https://arxiv.org/abs/1705.06950.

[17] M. Kim, H. B. Kim, J. Moon, J. Choi, and S. T. Kim. Do You Remember? Dense Video Captioning with Cross-Modal Memory Retrieval. In *CVPR*, 2024.

[18] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[19] X. Le, S. Manli, A. Anas, W. Jun, Y. An, P. Senthil, Z. Honglu, P. Viraj, D. Yutong, S. R. Michael, K. Shrikant, Z. Jieyu, Q. Can, Z. Shu, C. Chia-Chih, Y. Ning, T. Juntao, M. A. Tulika, H. Shelby, W. Huan, C. Yejin, S. Ludwig, C. Zeyuan, S. Silvio, C. N. Juan, X. Caiming, and X. Ran. xgen-mm (formerly blip-3): A family of open large multimodal models. *arXiv preprint*, 2024.

[20] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, pages 7492–7500, 2018.

[21] H. Mkhallati, A. Cioppa, S. Giancola, B. Ghanem, and M. V. Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries, 2023. URL https://arxiv.org/abs/2304.04565.

[22] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han. Streamlined dense video captioning. In *CVPR*, pages 6588–6597, 2019.

[23] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):1–15, July 2022. ISSN 1557-7368. doi: 10.1145/3528223.3530127. URL http://dx.doi.org/10.1145/3528223.3530127.

[24] T. Nitta, T. Fukuzawa, and T. Tamaki. Fine-grained length controllable video captioning with ordinal embeddings. *IEEE Access*, 12:123456–123467, 2024. doi: 10.1109/ACCESS.2024.3506751.

[25] M. Qi, Y. Wang, A. Li, and J. Luo. Sports video captioning via attentive motion representation and group relationship modeling. *TCSVT*, 30(8): 2617–2633, 2020. doi: 10.1109/TCSVT.2019.2921655.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1 (8):1–12, 2019.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*. IEEE, Jun. 2016.

[28] Y. Shi, X. Yang, H. Xu, C. Yuan, B. Li, W. Hu, and Z.-J. Zha. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In *CVPR*, pages 17929–17938, 2022.

[29] J. V. B. Soares, A. Shah, and T. Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *ICIP*, pages 2796–2800, Bordeaux, France, October 2022. Institute of Electrical and Electronics Engineers (IEEE).

[30] V. Somers, V. Joos, A. Cioppa, S. Giancola, S. A. Ghasemzadeh, F. Magera, B. Standaert, A. M. Mansourian, X. Zhou, S. Kasaei, et al. Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *CVPRW*, 2024. URL https://arxiv.org/abs/2404.11335.

[31] G. Suzuki, S. Takahashi, T. Ogawa, and M. Haseyama. Team tactics estimation in soccer videos based on a deep extreme learning machine and characteristics of the tactics. *IEEE Access*, 7:153238–153248, 2019.

[32] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, pages 7190–7198, 2018.

[33] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*. IEEE, Oct. 2021.

[34] W. Wu, H. Luo, B. Fang, J. Wang, and W. Ouyang. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *CVPR*, pages 10704–10713, 2023.

[35] W. Xu, J. Yu, Z. Miao, L. Wan, and Q. Ji. Bridging video and text: A two-step polishing transformer for video captioning. *TCSVT*, 32(9): 6293–6307, 2022.

[36] L. Yan, S. Ma, Q. Wang, Y. Chen, X. Zhang, A. Savakis, and D. Liu. Video captioning using global-local representation. *TCSVT*, 32(10): 6642–6656, 2022. doi: 10.1109/TCSVT.2022.3177320.

[37] L. Yan, Q. Wang, Y. Cui, F. Feng, X. Quan, X. Zhang, and D. Liu. GL-RG: Global-Local Representation Granularity for Video Captioning. In *IJCAI*, pages 2765–2771, 2022.

[38] L. Yan, C. Han, Z. Xu, D. Liu, and Q. Wang. Prompt learns prompt: Exploring knowledge-aware generative prompt collaboration for video captioning. In *IJCAI*, pages 1622–1630, 2023.

[39] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. *CoRR*, abs/2302.14115, 2023. URL http://arxiv.org/abs/2302.14115.

[40] D. Yang and C. Yuan. Hierarchical context encoding for events captioning in videos. In *ICIP*, pages 1288–1292, 2018.

[41] H. Yu, S. Cheng, B. Ni, M. Wang, J. Zhang, and X. Yang. Fine-grained video captioning for sports narrative. In *CVPR*, pages 6006–6015, June 2018.

[42] G. V. Zandycke and C. D. Vleeschouwer. 3d ball localization from a single calibrated image. In *CVPRW*, pages 3471–3479, New Orleans, LA, USA, June 2022. Institute of Electrical and Electronics Engineers (IEEE).

[43] Z. Zhang, D. Xu, W. Ouyang, and C. Tan. Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. *TCSVT*, 30(9):3130–3139, 2020. doi: 10.1109/TCSVT.2019.2936526.

[44] Y. Zheng, Y. Zhang, R. Feng, T. Zhang, and W. Fan. Stacked multimodal attention network for context-aware video captioning. *TCSVT*, 32(1): 31–42, 2022.

[45] L. Zhou, C. Xu, and J. Corso. Towards automatic learning of procedures from web instructional videos. *AAAI*, 32(1), Apr. 2018.

[46] X. Zhou, L. Kang, Z. Cheng, B. He, and J. Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv preprint arXiv:2106.14447*, 2021. URL https://arxiv.org/abs/2106.14447.

[47] X. Zhou, A. Arnab, C. Sun, and C. Schmid. Dense Video Object Captioning from Disjoint Supervision. *arXiv preprint arXiv:2306.11729*, 2023.

[48] H. Zhu, J. Liang, C. Lin, J. Zhang, and J. Hu. A transformer-based system for action spotting in soccer videos. In *MMSports*, pages 103–109, Lisbon, Portugal, October 2022. ACM.

[49] K. Zhu, A. Wong, and J. McPhee. Fencenet: Fine-grained footwork recognition in fencing. In *CVPRW*, pages 3588–3597, New Orleans, LA, USA, June 2022. Institute of Electrical and Electronics Engineers (IEEE).