Signal, Image and Video Processing (2025) 19:1024
https://doi.org/10.1007/s11760-025-04622-6

ORIGINAL PAPER

Real-time video captioning using feature fusion and attention mechanism

K R Sarath Chandran[1] · Dejah Madhusankar[1] · Krithika Swaminathan[1]

Received: 26 February 2025 / Revised: 23 July 2025 / Accepted: 4 August 2025

Abstract

The increasing prevalence and volume of short-form video content and real-time multimedia has made live video captioning indispensable. Consequently, there is a rising demand for autonomous systems that can understand the observed scene and describe it for applications ranging from large-scale content-indexing to live-commentary systems. Effective scene understanding requires precise object identification along with contextual interpretation of object interactions. This research achieves this through a novel feature fusion approach, combining low-level VGG16 CNN-extracted spatial features with YOLOv8 object detection results. The fused features are then processed through an LSTM-based encoder-decoder architecture, augmented with a customised category-aware attention mechanism, to enable real-time caption generation. Trained on the MSVD dataset, the proposed approach outperforms our baseline model by over 20% in BLEU-4 and METEOR scores, with a 54% drop in average inference time. Indeed, it is substantial to note that the motivation of this paper lies in improving foundational uni-modal methods for converting live video into natural language in rapid, resource-constrained environments, hypothesizing that advancements in this area can subsequently benefit the development of more efficient architectures leveraging lighter and more modern models.

Keywords Video captioning · Real-time · Feature extraction · Feature fusion · CNN · YOLOv8 · LSTM · Attention

1 Introduction

The shift from text-based to video-based content has transformed how we consume and communicate information. According to the Cisco Annual Internet Report (2023) [1], video content is expected to account for 82% of all internet traffic by 2024, driven largely by the rise of short-form content on platforms like TikTok, YouTube Shorts, and Instagram Reels. As visuals are processed faster and are more engaging than text [2, 3], video has emerged as the preferred medium of communication. However, the sheer volume of video content presents growing challenges in management, retrieval, and accessibility. Manual annotation is no longer feasible at scale, rendering automated video captioning essential for efficient indexing, enhancing search engine optimization, and improving accessibility. Real-time captioning further enables applications like live commentary, surveillance feed analysis, and assistive technologies. Thus, automatic scene understanding is a necessity.

Real-time video captioning demands rapid yet meaningful interpretation of visual data, which is difficult in constrained computational settings. Existing video captioning systems are often not optimized for real-time performance despite the existence of excellent pre-trained networks. Our approach aims to build on existing uni-modal architectures to enhance each part of the video captioning process and reduce latency without compromising accuracy.

To achieve this, this paper presents three key contributions. First, it introduces a feature fusion approach that combines low-level spatial features with objects detected from each set of frames. The core idea is that the superior object recognition provided by YOLOv8, when integrated

Sarath Chandran, Dejah Madhusankar and Krithika Swaminathan contributed equally to this work.

✉ K R Sarath Chandran
sarathchandran@ssn.edu.in

Dejah Madhusankar
dejah2010296@ssn.edu.in

Krithika Swaminathan
krithika2010039@ssn.edu.in

[1] Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai 603110, Tamil Nadu, India

Published online: 05 September 2025

with overall image features, enhances scene understanding by improving contextual correlation between objects in a scene. Second, it proposes a custom category-aware attention module that enables feature correlation through better contextual awareness while processing the varied features, facilitating improved caption generation. Third, it applies synonym substitution to the training labels during preprocessing to prioritize sentence accuracy over lexical diversity. The paper also demonstrates the effectiveness of its proposed approach by enabling relatively heavier modules to achieve reduced inference times with increased performance in a resource-constrained setup. It thereby lays the foundation for future integration with more modern, lightweight modules that may further boost both accuracy and speed. Together, these techniques aim to generate prompt, semantically rich captions suitable for real-time applications, with effectiveness validated through rigorous experimentation and analysis.

2 Related works

Recent research highlights the growing importance of real-time video captioning. For example, Zhou et al. [4] proposed a streaming model for dense video captioning that combined a fixed-size memory module with a streaming decoding algorithm to generate intermediate predictions before the entire video was processed, while Blanco-Fernández et al. [5] formally introduced the live video captioning task and defined a new evaluation metric to evaluate streaming systems. Liu et al. [6] established an Attentive Moment Retrieval Network (ACRN) that used a temporal memory attention mechanism to emphasize query-relevant visual features while incorporating context, producing an augmented representation of localized video segments. These works underscore the need for attention-based and low-latency approaches in modern video understanding systems, motivating our real-time captioning framework.

In addition to streaming concerns, many deep learning video captioning models focus on improving content understanding. For instance, Xu et al. [7] proposed a modality-specific LSTM (MA-LSTM) to isolate modality-relevant information, while Chen et al. [8] incorporated reinforcement learning to select only the most informative frames for caption generation. Olivastri et al. [9] explored end-to-end training that jointly optimizes encoding and decoding. On a slightly different trajectory, Zheng et al. [10] introduced a syntax-aware module to model relationships among video objects, while Hemalatha and Sekhar [11] employed domain-specific semantic identification. Fu et al. [12] used an NGPT kernel with Top-P sampling for effective topic-based caption generation. These approaches have shown strong results on standard datasets like MSVD.

Interestingly, recent research in image captioning approaches has inspired novel ideas on improving the feature selection process that we seek to extend to video captioning. Al-Malla et al. [13] presented an attention-based, encoder-decoder deep architecture that made use of convolutional features extracted from a CNN (Convolutional Neural Network) model pre-trained on ImageNet (Xception), along with object features extracted from the YOLOv4 (You Only Look Once) model pre-trained on the MS COCO (Microsoft Common Objects in Context) dataset. It is interesting that the output of these models were concatenated to give rise to a feature map that can be used to generate more accurate descriptions, a technique that improved CIDEr scores by 15% on the MS COCO and Flickr30k datasets. The work also introduced an importance factor for positional encoding of key objects, though this can now be more easily implemented with YOLOv8's confidence thresholding. Such fusion-based models, though primarily applied to image captioning, provide a promising foundation for improving video captioning, and our work extends this concept.

Further, the effectiveness of YOLO-inspired models themselves for captioning has been further supported in various studies. Karve et al. [14] proposed the YOLOv1-LSTM model, an improvised CNN that delivered faster output with good accuracy. Saroja and Mary [15] also achieved strong results using YOLOv5 with Xception V3 on the Flickr8k, Flickr30k and MS COCO datasets.

Within video captioning, Preethi and Dhanalakshmi [16] used pre-trained CNNs (Inception V3, VGG16) with LSTM for caption generation, though the system suffered from slow inference and object detection inaccuracies. Chougole and Chavan [17] proposed a hybrid CNN-FFT approach with BiLSTM to capture both spatial and temporal features, showing strong performance on MSR-VTT. Agarwal et al. [18] combined YOLOv8-based object detection with visual features to caption apparel images and videos effectively, demonstrating the extensibility of such fusion methods for scene understanding. This is further supported by the effectiveness of the CNN-LSTM combination for video surveillance, as evidenced in the studies on anomaly detection done by Amin et al. [19–21], justifying our choice for the foundation of the proposed architecture. Such an application confirms the robustness of the CNN-LSTM combination in identifying events and temporal patterns in streaming footage.

Despite the breadth of work, two key gaps remain: 1) Real-time readiness is often compromised in favor of deeper comprehension, making many systems impractical for low-resource or latency-sensitive environments. 2) fusion of global and object-level cues remains underexplored in video captioning, despite its success in image captioning. Few approaches attempt to optimize the two parts of the captioning process individually and later combine them to incorporate the best of both tasks.

Our system builds on these gaps. Unlike prior models, it is engineered for real-time performance on modest hardware by reducing redundant computations (via fused features instead of multiple backbones) and employing tailored attention mechanisms to fully exploit the combined representation. In this way, our work fills the identified niche: delivering a unified, real-time video captioning architecture that that balances speed, accuracy, and computational cost.

3 Proposed system

We present two variations, **CYF-KAL** (CNN-YOLOv8 Fusion with Keras Attention Layer) and **CYF-CAL** (CNN-YOLOv8 Fusion with Custom Attention Layer), of our proposed system for real-time scene identification and description. Both models are trained on the complete MSVD dataset using Google Colaboratory's [22] v2-8 TPU infrastructure with 343 GB of RAM. The system architecture is illustrated in Figure 1, with the data flow and underlying design rationale discussed in the following paragraphs.

A video input to the system undergoes spatial and object-oriented feature extraction, followed by scene understanding using an attention-augmented LSTM-based encoder and generation of a textual description using a decoder module. For extended video sequences, the input is taken one segment at a time, where each segment is of a duration fixed based on the user's requirements, with the system generating a single caption for each segment. For a live stream such as a monitoring feed with scene changes expected at a less frequent rate, a bigger segment size can be chosen to reduce redundant processing. Similarly, a live stream with rapid scene changes can be set to be processed in much smaller segments, say, 3-6 seconds, for increased information capturing and caption specificity.

Starting with the visual model in Figure 1, the system processes each video segment by uniformly sampling a predefined number of frames. To ensure comprehensive scene understanding, we employ a dual-stream feature extraction approach that captures both spatial features–such as shapes, textures, and contours–and object-level information from each frame, hypothesizing that low-level spatial cues can facilitate the identification of interactions between detected objects. For this, we leverage two pre-trained models: VGG16 (Visual Geometry Group) CNN [23] (on ImageNet [24]) for spatial features and YOLOv8n [25] (on MS-COCO [26]) for object detection.

VGG16 is selected for its stable feature representations, ease of integration, and widespread use in prior captioning systems, which enables fair benchmarking. Its uniform convolutional architecture also ensures stable per-frame latency, avoiding runtime jitter in live settings. Importantly, as a relatively heavy model by modern standards, its computational demands emphasize the substantial inference latency reductions achieved by our architectural contributions, which constitute the core novelty of our work. Similarly, YOLOv8 presents itself as suitable due to it being the first in its series to combine state-of-the-art detection with per-object confidence scores, enhancing contextual modeling. We adopt the YOLOv8n variant, the fastest and most lightweight in the family, to ensure minimal inference overhead during frame-wise object detection and to support real-time responsiveness. YOLOv8n also offers sufficient accuracy, rendering larger variants as over-provisioned for our needs. We utilize the penultimate fully connected layer of VGG16 and the final detection output of YOLOv8 to create a fused feature vector representing each video segment. Thus, both VGG16 and YOLOv8 are used as frozen, off-the-shelf feature extractors, with weights intact during training.

The fused features are then passed on to a sequence-to-sequence architecture built with LSTM (Long Short-Term Memory) [27] cells. The encoder component imposes the detected object information on the extracted spatial features and uses them to derive better temporal inferences across the multiple frames in a segment. To support the feature fusion, a trainable attention layer is also introduced, which is further elaborated in Section 2 of Online Resource 1. The decoder component, trained on the context vector from the encoder (refer Section 3 of Online Resource 1) and the corresponding training label, generates a suitable caption for each video segment in real-time as outlined in Section 4 of Online Resource 1.

This architectural choice is driven by practical trade-offs. LSTMs effectively model long-term dependencies in sequential data, allowing for temporal coherence, which is a key requirement in video captioning, with low latency and memory usage, making them suitable for real-time applications under resource constraints. Unlike Transformer-based models such as T5 [28] and SwinBERT [29], which require more compute and use fixed self-attention, LSTMs within an encoder-decoder setup allow for flexible, customizable attention mechanisms, such as the category-aware attention used here, enabling fine-grained control over feature weighting and interpretability.

The novelty of this design is perhaps more incremental and practical than a groundbreaking theoretical advancement, but significant nonetheless. While the fusion of CNN and YOLO features has been explored in image captioning, this system applies a structured, fine-grained fusion approach specifically tailored for real-time video scenarios, with an emphasis on reducing caption generation time. Such a strategy remains relatively underexplored for temporal captioning. Moreover, the attention mechanism used here is category-aware, unlike standard approaches (Luong, Bahdanau) which treat input

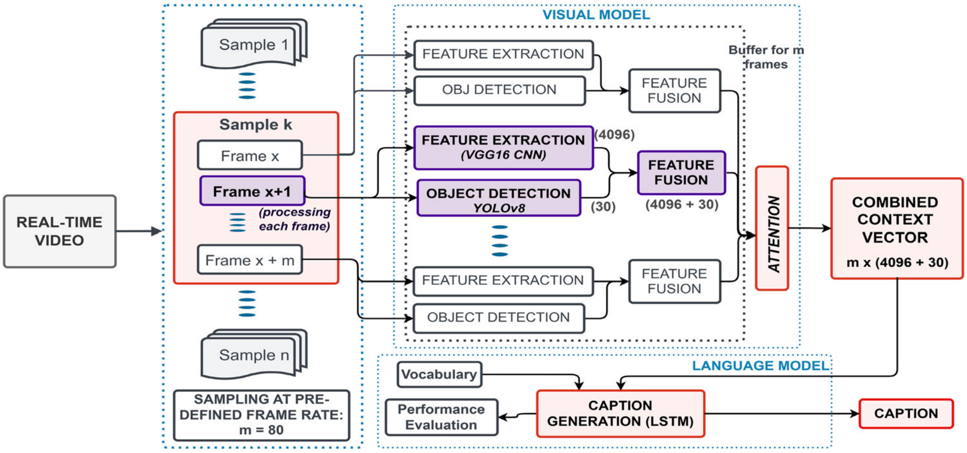


Fig. 1 Architecture of the Proposed System

features as flat vectors, allowing for more interpretable and effective feature weighting.

4 Dataset and preprocessing

The Microsoft Research Video Description Corpus (MSVD) [30] is a collection of 1970 videos, each video depicting a single activity that lasts for about 6 to 25 seconds. For example, the video may be a small clip of someone cooking, of a man riding a bike, or of a baby playing. Each video clip has multiple associated captions across multiple languages, featuring over 120,000 sentences with an average of 15 English captions per clip. Following the evaluation setup of prior work [31], we split the dataset into three parts: 1,200 videos for training, 100 for validation, and 670 clips for testing.

On observing the captions in the dataset, we notice that the frequency of important words in the given captions is needlessly reduced due to some characteristics of the annotations. For instance, synonyms foster variation in word choice. Therefore, to ensure consistency, we prioritize frequent words and replace all synonyms. Some examples are shown in Table 1.

Only the refined captions of sentences of length 6–10 words are retained. They are tokenized and padded to a uniform length of 10 words. Each caption phrase is wrapped within the delimiters ⟨ bos⟩ and ⟨ eos⟩. A Keras Tokenizer [32] is used to create a vocabulary list comprising a unique set of the most frequently used words in the captions. The top 3700 (corroborated in Section 5.1) words are identified from this list to serve as the vocabulary for the language model.

5 Experiments

5.1 Ablation study

To quantify the contribution of each component in our proposed model, we conduct an ablation study by systematically altering or removing individual modules. An analysis of parameter variations is also conducted to empirically substantiate that the proposed model architecture is constructed using an optimal configuration.

Experiments for this study are conducted on the NVIDIA Tesla T4 GPU with 16GB of VRAM provided by Google Colaboratory. To manage high computational requirements, a reduced subset of the MSVD dataset comprising 755 training and 416 test videos is used, preserving the original train-to-test ratio (1200:670::755:416). The data preprocessing steps remain consistent, applying caption length filtering and synonym substitution as outlined in Section 4. A new vocabulary of 1500 tokens is, however, extracted from this subset and retained across the study, except in the fourth variation, where vocabulary size is explicitly modified for experimentation. To ensure a proportional scale-down, the batch size is adjusted accordingly to 380 (as per the ratio 1200:755::600:380). Model performance is evaluated using the metrics detailed in Section 1 of Online Resource 2, with comprehensive results presented in Table 2.

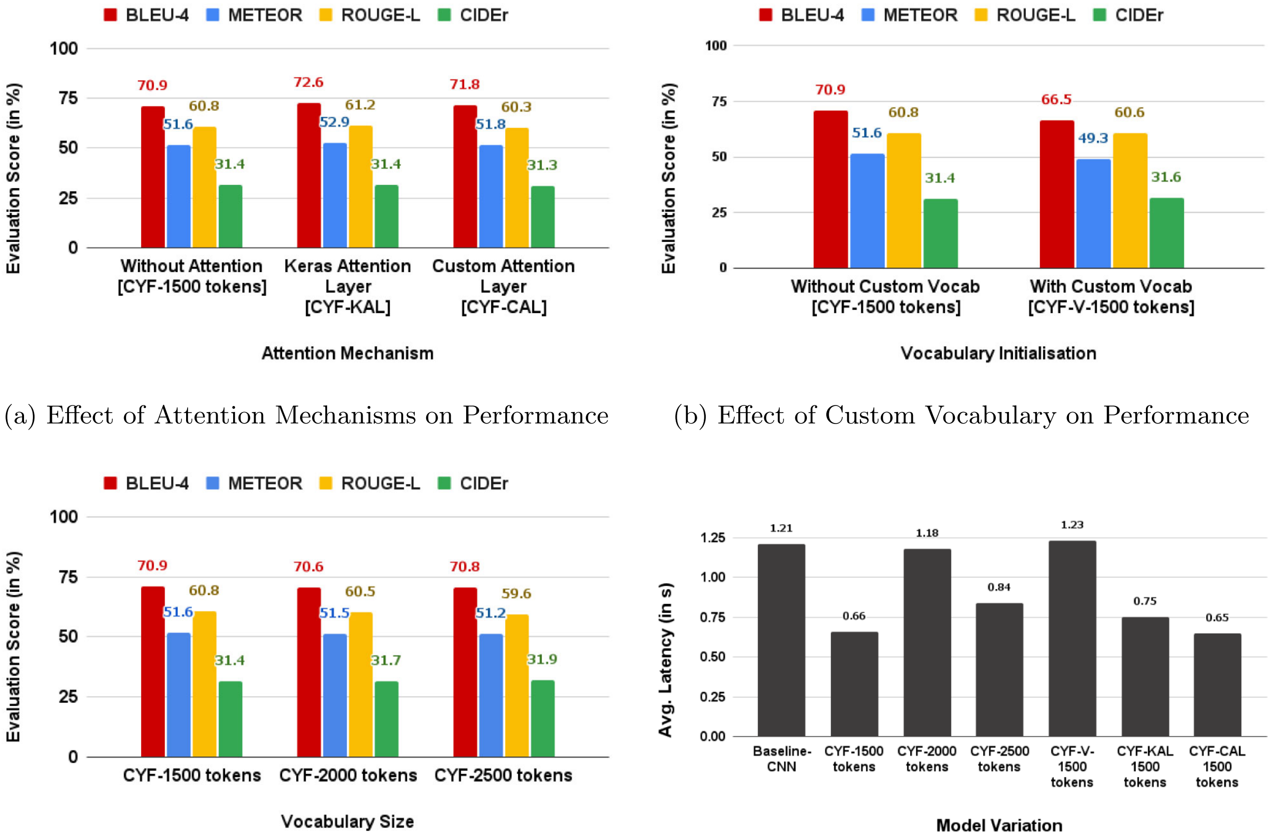
1. **Varying Attention Module:** We begin the study by building the scaled-down variants of our proposed models, in the environment specified for this study, to ensure

Table 1 Examples for Replacing Synonyms

Synonyms	Replaced by
infant, newborn	baby
huge, large	big
individual, lad, man/woman, male/ female, lady	person
many, a number of	several

Table 2 Ablation Study; EPOCHS:Early Stopping out of 120 epochs, T-A:Training Accuracy(%), V-A:Validation Accuracy(%), B:BLEU-4(%), M:METEOR(%), R:ROUGE-L(%), C:CIDEr(%), T:Average inference time for caption generation (in seconds)

MODEL	EPOCHS	T-A	V-A	B	M	R	C	T
Fusion: CNN + YOLOv8 + Attention + LSTM Enc-Dec with 1500 tokens								
Keras Attention layer	97	81.14	73.72	72.6	52.9	61.2	31.4	0.75
Custom Attention layer	78	79.05	74.26	71.8	51.8	60.3	31.3	0.65
Fusion: CNN + YOLOv8 + Custom Vocabulary + LSTM Enc-Dec								
1500 tokens	74	78.78	73.52	66.5	49.3	60.6	31.6	1.23
Fusion: CNN + YOLOv8 + LSTM Enc-Dec								
1500 tokens	91	80.88	73.49	70.9	51.6	60.8	31.4	0.66
2000 tokens	73	77.83	73.02	70.6	51.5	60.5	31.7	1.18
2500 tokens	66	75.78	71.70	70.8	51.2	59.6	31.9	0.84
Baseline: CNN + LSTM Enc-Dec								
1500 tokens	70	75.87	70.79	53.1	33.4	50.2	27.6	1.21



(a) Effect of Attention Mechanisms on Performance

(b) Effect of Custom Vocabulary on Performance

(c) Effect of Vocabulary Size on Performance

(d) Avg. Inference Time (secs) across Model Variations

Fig. 2 Comparison of Model Performance

a fair analysis. The impact of the two attention mechanisms, elaborated in Section 2 of Online Resource 1, on caption generation performance is observed in isolation. Figure 2a clearly suggests that both techniques perform similarly (differing by 1-2% at most), with the built-in Keras-API attention layer only achieving marginally higher BLEU-4 (72.6%) and METEOR (52.9%) scores. While the Keras model exhibits higher training accuracy (81.14%), the custom variant converges faster (78 vs. 97 epochs) with better validation accuracy (74.26% vs. 73.72%). This suggests that the Keras variant may overfit slightly, whereas the Custom attention promotes better generalization, likely due to a effective regularization.

2. **Removing Attention Module:** To quantify the essentiality of the attention mechanism itself, we remove it entirely and feed the fused feature set directly into the LSTM encoder. However, Figure 2a shows that this results in almost a 2% drop in BLEU-4, with a consistent degradation across all metrics in the non-attention variant (CNN + YOLOv8 + LSTM) as per Table 2. This affirms that the addition of attention significantly enhances focus and information correlation while learning.
3. **Augmenting Vocabulary Set** To improve caption quality, we experiment with augmenting the tokenized vocabulary by prefixing a curated set of common English words - including articles *(a, an, the)*, gerunds, popular nouns (cross-referenced with MS-COCO class labels), and number words - to the base vocabulary extracted from training captions (Section 4). However, as shown in Figure 2b , this augmentation yields no performance gain. On the contrary, BLEU-4 drops to 66.5%, and CIDEr falls to 31.6%, with caption generation latency nearly doubling (1.23s vs. 0.66s). The results suggest that manually injecting vocabulary introduces lexical biases and sparsity, degrading model efficiency and generalization. Larger token spaces are also harder to learn from in limited data settings. Consequently, this strategy is discarded in favor of the original token set.
4. **Varying Vocabulary Size:** After removing the custom vocabulary, we investigate the impact of increasing token size, hypothesizing that a broader vocabulary may enhance caption specificity. Token counts are varied from 1500 in increments of 500. However, as shown in Figure 2c , performance declines with larger vocabularies. Increased token sets introduce rare or irrelevant words, making the language more prone to outliers, degrading sentence structure and coherence. We also see caption generation latency rise notably (1.18s at 2000 tokens vs. 0.66s at 1500), while early training convergence at 66 epochs indicates diminished learning capacity. The expanded softmax space likely hampers decoder efficiency, further slowing inference In contrast, the 1500-token model achieves superior metrics, offering a favorable trade-off between performance and computational cost. Accordingly, we standardize 1500 as the optimal vocabulary size for our model.
5. **Removing Fusion of YOLOv8 Features:** Taking its place as perhaps the most important study to underscore the role of the feature fusion process, we define our baseline [33] for video captioning as the minimal encoder-decoder architecture using CNN-only spatial features without object detection, attention mechanisms or synonym-substitution layers. Table 2 shows that the baseline exhibits significantly lower performance than our fusion-based variations, particularly in METEOR scores, which are sensitive to semantic accuracy, stemming, and synonym handling; highlighting the relevance of such a fusion (described in detail in Section 1 of Online Resource 1). Interestingly, despite its architectural simplicity, the baseline also incurs higher latency (1.21s) compared to the YOLOv8+LSTM fusion model (0.66s), indicating a 50% reduction in inference time. This counterintuitive result suggests that YOLOv8 provides compact, semantically rich features that alleviate decoder load, thereby improving both speed and output quality. Moreover, the pronounced decline in BLEU-4 also underscores the contribution of our synonym-substitution strategy (see Section 4) to improved accuracy. Conceivably, the susbtitution reduces the model's redundant processing of synonymous terms while maintaining semantic relevance in generated captions.

Analysis of Table 2 confirms that the full model - incorporating YOLOv8 features, attention mechanisms, and a well-calibrated vocabulary - consistently outperforms all other variations across all metrics. Notably, despite the inclusion of more complex modules, captioning latency significantly improves from 1.21s (baseline) to 0.65s (full model with custom attention layer), denoting a near 50% reduction in time. This substantial latency reduction is critical for real-time applications, as illustrated in Figure 2d .

Given the complementary nature of all components, we thus justify the architecture adopted in our proposed model. To ensure consistent performance and allow for a robust corroboration from the ablation study, the parameters are ensured to be scaled proportionally. A Python script analyzing the complete MSVD caption set reveals 36,502 unique words, compared to 14,634 in the subset of 755 samples. Accordingly, we adjust the top N vocabulary words utilized by our proposed models, resulting in a final vocabulary size of 3,700 tokens (36502:14634::3700:1500). A second layer of verification, carried out by maintaining the vocabulary size as 1500 tokens while training the proposed models, resulted in an approximate 2% decrease across all metrics. Thus,

Table 3 Comparison of average inference time (latency) in caption generation across different computing environments

Model	Inference Time (s)		
	CPU	GPU	TPU
CYF-KAL	0.78	0.73	0.59
CYF-CAL	0.78	0.71	0.57

CYF-CAL and CYF-KAL adopt a vocabulary size of 3700, showing consistency with our calculations.

5.2 Hardware-aware latency study

To evaluate the efficiency of caption generation across different computing environments, we benchmark our final proposed models, CYF-CAL and CYF-KAL, on hardware configurations available via Google Colaboratory: CPU, GPU, and TPU (v2-8) (see Section 2 of Online Resource 2 for detailed hardware specifications). The MSVD test set, as described in Section 4, is used for evaluation, and Table 3 presents the average inference time per test video across the hardware platforms.

As anticipated, both models exhibit a progressive reduction in latency from CPU to GPU to TPU due to increased computational parallelism. CYF-CAL consistently demonstrates faster inference times compared to CYF-KAL, achieving approximately 2.7% reduction in latency on GPU and 3.4% on TPU. On CPU, both models show identical higher inference times (0.78s), indicating that the efficiency gains of the custom attention layer are realized primarily through hardware accelerators capable of parallelized tensor operations. This makes the architecture better suited for real-time or latency-sensitive applications, especially those deployed on edge devices or cloud-based platforms. Additionally, caption outputs remain consistent across platforms, confirming hardware independence in performance quality.

6 Results and analysis

The two proposed models, CYF-KAL and CYF-CAL (refer Section 3, are trained on the entire MSVD dataset, and the observed results are presented in Table 4. The subsequent sections provide a detailed analysis of the same.

6.1 Comparison with SOTA models

Table 5 presents a performance comparison between the proposed models and several state-of-the-art approaches on the MSVD dataset. Our models achieve superior BLEU-4 and METEOR scores, indicating strong lexical and grammatical alignment with reference captions, while exhibiting lower ROUGE-L and CIDEr scores. This suggests that our integration of both spatial and object-level features, along with synonym substitution in captions and an optimized vocabulary size, enables precise frame understanding and word selection while potentially hindering semantic understanding due to a restricted vocabulary.

Analysis of Table 5 reveals certain key findings. PickNet [8] demonstrates the viability of selective frame-based video captioning by selecting a sparse set of frames through reinforcement learning to reduce redundancy. Meanwhile, our approach samples frames at regular time intervals to ensure that the entire video stream is represented, producing captions that better summarize the video content. Although this results in a lower CIDEr score, CYF-KAL and CYF-CAL surpass PickNet on the other three metrics.

Arguably, the two-stage approach of first optimizing individual components before fine-tuning the entire network appears to reduce inference times over end-to-end training schemes such as those of the EtENet models [9]. Although the lower ROUGE-L and CIDEr scores indicate that the generated captions may be less semantically diverse, the balance between speed and accuracy discovered in the proposed model is crucial in dynamic environments where immediate caption generation is necessary.

Meanwhile, DSD-3 DS-SEM [11] and SAAT [10] prioritize domain-specific semantics and sentence syntax, respectively, potentially explaining their dominance in the ROUGE-L and CIDEr scores. In addition to these, recent research has further advanced semantic modeling in video captioning. NGPT-Neo [12] introduces a semantic topic association method that integrates object and action-level topic extraction with a GPT-based decoder enhanced by Top-P nucleus sampling. This approach achieves significant improvements on the MSVD dataset, including a 13.1% increase in CIDEr scores, by effectively narrowing the semantic gap between video content and generated captions.

Further extending the state of the art, GDKRL [34] employs graph-based deep knowledge representation learning to model complex inter-object relationships within video frames. Its performance on MSVD demonstrates high scores across all metrics while maintaining an efficient inference time of 112 ms, highlighting its suitability for near real-time applications. GL-RG [35] contributes a granularity-aware visual representation strategy, capturing both global and local video context, which improves linguistic expressiveness in generated captions. This is showcased in its high ROUGE-L and CIDEr scores. TextKG [36], while taking a multimodal approach and using a cross-attention mechanism to share information across streams, enhances transformer-based captioning by integrating external knowledge graphs to resolve long-tail word challenges, thereby significantly boosting CIDEr scores. Similarly, VSLAN [37] leverages variational stacked local attention mechanisms to enrich cap-

Table 4 Performance of Proposed Models on Complete MSVD Dataset; EPOCHS:Early Stopping out of 120 epochs, T-A:Training Accuracy(%), V-A:Validation Accuracy(%), B:BLEU-4(%), M:METEOR(%), R:ROUGE-L(%), C:CIDEr(%), T:Average inference time for caption generation (in seconds). CYF-KAL: CNN-YOLOv8 Fusion with Keras Attention Layer, CYF-CAL: CNN-YOLOv8 Fusion with Custom Attention Layer

PROPOSED MODEL	EPOCHS	T-A	V-A	B	M	R	C	T
CYF-KAL	40	69.25	62.44	74.40	54.30	63.40	33.30	0.73
CYF-CAL	43	69.54	62.55	73.50	53.40	62.40	32.70	0.71

Table 5 Performance comparison with state-of-the-art approaches. B:BLEU-4(%), M-METEOR(%), R:ROUGE-L(%), C:CIDEr(%). The two best scores with respect to each metric except CIDEr are highlighted in bold

APPROACH	B	M	R	C
MA-LSTM [7]	36.5	26.5	41	59.8
PickNet [8]	41.3	27.7	44.1	59.8
DSD-3 DS-SEM [11]	50.1	34.7	**73.1**	76
SAAT [10]	46.5	33.5	69.4	81
EtENet-IRv2step1 [9]	49.1	33.6	69.5	83.5
EtENet-IRv2step2 [9]	50	34.3	70.2	60.6
NGPT-Neo [12]	58.8	38.9	75.1	103.2*
GDKRL [34]	**77.9**	39.2	74.9	101.3*
GL-RG [35]	60.5	38.9	**76.4**	101*
TextKG [36]	60.8	38.5	75.1	105.2*
VSLAN [37]	57.4	36.9	**75.6**	98.1
MAN [38]	59.7	37.3	74.3	101.5*
CYF-KAL (proposed approach)	**74.4**	**54.3**	63.4	33.3
CYF-CAL (proposed approach)	73.5	**53.4**	62.4	32.7

*In cases where original CIDEr scores from other studies exceed 100%, they are retained as reported, reflecting the unnormalized scoring convention

tion diversity through deeper multimodal interaction. Finally, MAN [38] introduces a memory-augmented architecture that incorporates implicit external knowledge during decoding, resulting in more contextually grounded captions.

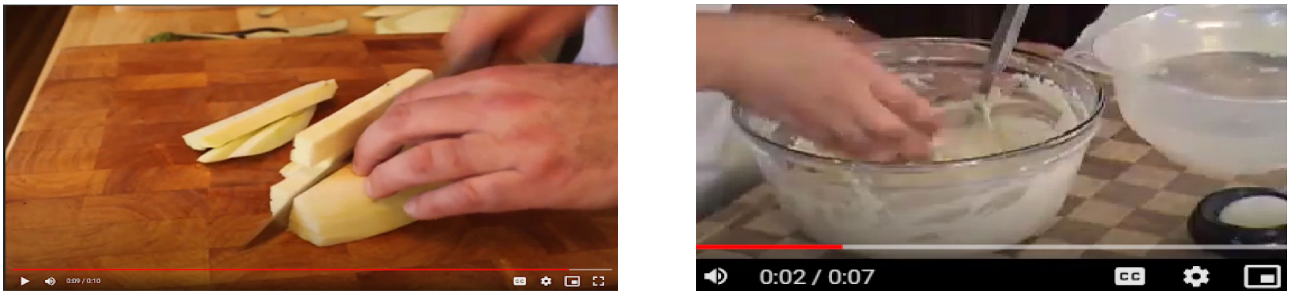
With better computational power aiding the maintenance of a dynamic vocabulary or further refinement of the attention mechanisms, we anticipate further gains in semantic richness (as captured by ROUGE-L and CIDEr) without compromising the low latency or lexical precision of our proposed model.

6.2 Qualitative analysis

Both variations of our proposed model generate high-quality captions with subtle differences, as shown in Figures 3a and 3b . The generated outputs are analyzed on the basis of the interplay between object specificity and generalization, contextual coherence, and the implications of the applied attention mechanism. For instance, in Figure 3a , CYF-KAL generates the caption "*a person is slicing a potato*," demonstrating a good degree of object specificity. The delimited identification of the object class indicates that the Luong-style attention mechanism effectively integrates the object detection output of YOLOv8 into the caption generation process. CYF-KAL also shows capabilities of capturing spatial and sequential details with "*a person is putting ingredients in a bowl*" as seen in Figure 3b . This explicit encoding of step-wise actions on objects is likely due to direct weighting on YOLOv8's detected objects and positional context, leading to highly specific, temporally grounded action descriptions. However, this sensitivity to detail also increases susceptibility to object detection errors in the early stage propagating into the caption.

Conversely, CYF-CAL generates the caption *"a person is slicing a vegetable"* for Figure 3a , exhibiting a shift toward generalization by replacing the specific term "potato" with the broader term "vegetable" while still maintaining contextual relevance. This trend is consistent across different image scenarios where, in Figure 3b , CYF-CAL generates "*a person is mixing ingredients*", again opting for a broader description of the action. While the custom attention layer preserves sentence structure, it appears to distribute attention more broadly across the input features, leading to a less confident classification of the specific object. This cautious approach of avoiding over-reliance on object detection output enhances robustness against misclassification but sacrifices fine-grained detail - a generalization that may be advantageous when the detection confidence of YOLOv8 is low, mitigating the risk of incorrect object labeling. This is especially reiterated when the caption generation model is reliant on a fixed vocabulary, furthering the propensity for mislabeling.

Analysis of the caption generation times reveals interesting trends. While both models exhibit comparable inference times for the video in Figure 3a (0.64s for both CYF-KAL and CYF-CAL), a notable difference emerges in the second scenario. CYF-KAL requires 0.71s to generate its more detailed caption, while CYF-CAL generates its more abstract description in just 0.50s. This suggests that the computational cost is not solely determined by the model architecture, but also influenced by the level of specificity captured in the gener-



(a) CYF-KAL: a person is slicing a potato; CYF-CAL: a person is slicing a vegetable

(b) CYF-KAL: a person is putting ingredients in a bowl; CYF-CAL: a person is mixing ingredients

Fig. 3 Caption generated for a sample video by variations of the proposed model

ated caption. This allows us to favor CYF-CAL for use-cases that rely on speedy results with less room for errors.

Shifting focus to the naturalness of formed sentences, both models generate grammatically correct and coherent captions. Notably, the shared phrase *"a person is slicing"* from both models indicates effective encoding of human-object interaction. However, it is relevant to mention here that while the refined vocabulary we have experimented with is less complex and gender neutral, the limited vocabulary may restrict scene understanding and occasionally omit crucial sentence elements. Additionally, object misidentification (for example, person as baby, or vice versa) may sometimes occur due to feature similarities or ground-truth annotation ambiguities.

7 Conclusion and future work

This research proposes a novel real-time video captioning approach that performs uni-modal feature fusion of spatial and object-level information, combined with a custom modelled category-aware attention-integrated LSTM for enhanced scene understanding and caption generation. Trained on the MSVD dataset, the system offers significantly reduced latency, improved lexical performance, and efficient resource usage – key strengths that support its suitability in adapting to real-time applications.

However, certain limitations remain. The study operates with a restricted vocabulary and lacks cross-dataset validation, with occasional semantic lapses arising from the limited diversity of training data. Additionally, the system remains primarily tested in simulated settings rather than actual real-world scenarios. Nonetheless, the significant improvements introduced by our approach in latency and caption accuracy imply that extending the methodology to an architecture leveraging more sophisticated and lightweight modules could drastically improve performance. Future work will thus explore replacing VGG16 with architectures such as EfficientNetV2, MobileNetV3 for faster inference, or YOLOv8 with the latest YOLO variants. Refinement strategies may help improve noun-level caption specificity, while boundary detection or frame comparison can reduce redundancy by captioning only semantically distinct scenes. Finally, Transformer-based variants may be investigated to further enhance semantic richness and long-range coherence, provided computational demands remain feasible for real-time deployment. Indeed, with further improvements, the approach explored here could be extended to real-time scene understanding in assistive technologies, video indexing, and surveillance applications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11760-025-04622-6.

Author Contributions Dr. K. R. Sarath Chandran contributed to the idea formulation and conceptualization of the study, research design and methodology planning, and overall project coordination and management. He was also responsible for editing, revision, and finalization of the manuscript, as well as presentation and outreach. Ms. Dejah was involved in research design and methodology planning, dataset research, and data curation, including tool selection for the study. She implemented the baseline models, conducted model training with an attention mechanism, and performed an in-depth analysis of the results. Additionally, she contributed to figure creation, manuscript drafting, and paper writing, along with editing, revision, and finalization. Ms. Krithika worked on research design and methodology planning, conducted an extensive literature review, and performed data preprocessing, feature extraction, and fusion. She trained model variations, conducted testing and evaluation, and compared performance with state-of-the-art methods. She also contributed to manuscript drafting, and paper writing, as well as editing, revision, and finalization. All authors reviewed and approved the final version of the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Cisco: Cisco Report. https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html (2023)
2. Pictory: The Power of Visual Communication: Why Video Brings Text to Life. https://pictory.ai/blog/the-power-of-visual-communication-why-video-brings-text-to-life (n.d.)
3. Granitz, N., Kohli, C., Lancellotti, M.P.: Textbooks for the YouTube generation? a case study on the shift from text to video. Journal. Educa. Bus. **96**(5), 299–307 (2021)
4. Zhou, X., Arnab, A., Buch, S., Yan, S., Myers, A., Xiong, X., Nagrani, A., Schmid, C.: Streaming dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18243–18252 (2024)
5. Blanco-Fernández, E., Gutiérrez-Álvarez, C., Nasri, N., Maldonado-Bascón, S., López-Sastre, R.J.: Live video captioning. Multimedia Tools and Applications, 1–33 (2025)
6. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.-S.: Attentive moment retrieval in videos. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 15–24 (2018)
7. Xu, J., Yao, T., Zhang, Y., Mei, T.: Learning multimodal attention LSTM networks for video captioning. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 537–545 (2017)
8. Chen, Y., Wang, S., Zhang, W., Huang, Q.: Less is more: Picking informative frames for video captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 358–373 (2018)
9. Olivastri, S., Singh, G., Cuzzolin, F.: End-to-end video captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
10. Zheng, Q., Wang, C., Tao, D.: Syntax-aware action targeting for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13096–13105 (2020)
11. Shekhar, C.: Domain-specific semantics guided approach to video captioning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1587–1596 (2020)
12. Fu, Y., Yang, Y., Ye, O.: Video captioning method based on semantic topic association. Electronics **14**(5), 905 (2025)
13. Al-Malla, M.A., Jafar, A., Ghneim, N.: Image captioning model using attention and object features to mimic human image understanding. J. Big. Data. **9**, 20 (2022)
14. Karve, P., Thorat, S., Mistary, P., Belote, O.: Conversational Image Captioning Using LSTM and YOLO for Visually Impaired. In: Proceedings of Third International Conference on Communication, Computing and Electronics Systems, vol. 844, pp. 851–862. Springer, Cham (2022)
15. Saroja, M., Mary, A.B.: Image Captioning Using Improved YOLO V5 Model and Xception V3 Model (2023)
16. Preethi, A., Dhanalakshmi, P.: Video Captioning using Pre-Trained CNN and LSTM. In: 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), pp. 1–7 (2023)
17. Chougule, A.R., Chavan, S.D.: Custom cnn-bilstm model for video captioning. Multimedia Tools and Applications, 1–26 (2024)
18. Agarwal, G., Jindal, K., Chowdhury, A., Singh, V.K., Pal, A.: Image and video captioning for apparels using deep learning. IEEE Access 12, 113138–113150 (2024) https://doi.org/10.1109/ACCESS.2024.3443422
19. Amin, S.U., Ullah, M., Sajjad, M., Cheikh, F.A., Hijji, M., Hijji, A., et al.: Eadn: An efficient deep learning model for anomaly detection in videos. mathematics 2022 (2022)
20. Ul Amin, S., Kim, Y., Sami, I., Park, S., Seo, S.: An efficient attention-based strategy for anomaly detection in surveillance video. Computer Systems Science & Engineering **46**(3) (2023)
21. Ul Amin, S., Kim, B., Jung, Y., Seo, S., Park, S.: Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3d convolutions and long short-term memory modules. Adv. Intell. Syst. **6**(7), 2300706 (2024)
22. Google: Google Colaboratory. https://colab.google/
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014). https://arxiv.org/abs/1409.1556
24. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
25. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Version 8.0.0). https://github.com/ultralytics/ultralytics (2023)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. Springer (2014)
27. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
28. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
29. Lin, K., Li, L., Lin, C.-C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L.: Swinbert: End-to-end transformers with sparse attention for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17949–17958 (2022)
30. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 190–200 (2011)
31. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)
32. Keras: Keras Tokenizer. https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer (n.d.)
33. GitHub: Video-Captioning. https://github.com/Shreyz-max/Video-Captioning
34. Shankar, M.G., Surendran, D.: An effective video captioning based on language description using a novel graylag deep kookaburra reinforcement learning. EURASIP J. on Image and Video Process. **2025**(1), 1 (2025)
35. Yan, L., Wang, Q., Cui, Y., Feng, F., Quan, X., Zhang, X., Liu, D.: Gl-rg: Global-local representation granularity for video captioning. arXiv preprint arXiv:2205.10706 (2022)
36. Gu, X., Chen, G., Wang, Y., Zhang, L., Luo, T., Wen, L.: Text with knowledge graph augmented transformer for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18941–18951 (2023)

37. Deb, T., Sadmanee, A., Bhaumik, K.K., Ali, A.A., Amin, M.A., Rahman, A.: Variational stacked local attention networks for diverse video captioning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4070–4079 (2022)
38. Jing, S., Zhang, H., Zeng, P., Gao, L., Song, J., Shen, H.T.: Memory-based augmentation network for video captioning. IEEE Trans. Multimedia **26**, 2367–2379 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.