Video captioning of object related activities using estimation of object shift vectors

Prashant Kaushik
Department of CSE&IT
Jaypee Institute of Information Technology, Noida
jiitprashant@gmail.com

Vikas Saxena
Department of CSE&IT
Jaypee Institute of Information Technology, Noida
vikas.saxena@jiit.ac.in

Amarjeet Prajapati
Department of CSE&IT
Jaypee Institute of Information Technology, Noida
amarjeetnitkkr@gmail.com

Abstract: Object-oriented video captioning focuses on generating descriptions, summarization that emphasize object-related activities, including basic interactions such as object-object dynamics their interection with background and engagement with the background. Traditional methods primarily provide general video descriptions by extracting spatial and temporal features but often fail to capture how objects interact. To address this gap, our study examines object movements and trajectories to estimate interactions within a scene. We propose an innovative approach that tracks object motion and activities based on their paths, extracting valuable features that combine movement and interaction patterns. These features—such as grouped activities and collisions—are then used to train a transformer model, enabling context-aware, object-focused text generation. Additionally, we have developed a specialized object-oriented dataset to improve the quality of interaction-centered captions. Our approach is evaluated on subsets of two widely recognized datasets, MSVD and MSR-VTT, comparing the generated captions with state-of-the-art methods both quantitatively and qualitatively, demonstrating the advantages of our methodology.

Keywords: Geometric transformations, monitoring several objects, periodic motion, video captioning, estimating speed, and video captioning, objected oriented video captioning, video understanding.

1. Introduction

Video captioning has become an essential tool in various domains, including drone surveillance, autonomous systems, and traffic monitoring. These techniques are particularly effective in identifying anomalies within video footage, such as detecting similar sequences and recognizing irregular activities like wrong-way driving. By automating these processes, video captioning significantly reduces the reliance on manual supervision, saving time and human effort. However, as the field advances, there is an increasing demand for more sophisticated captioning techniques that focus on object motion and interaction their collision and background movements, leading to the emergence of object-oriented video captioning. [1][6]

Object-oriented video captioning aims to identify objects within a scene, analyze their movements and interactions, collision, and generate textual descriptions that accurately represent these dynamics. Object motion plays a crucial role in understanding scene context, as movement patterns reveal interaction types, intent, and behaviour .. However, conventional and deep learning-based video captioning methods often prioritize scene-wide spatial and temporal features while overlooking detailed object motion and inter-object interactions This limitation results in captions that lack depth in describing object movements, leading to incomplete or less meaningful narratives.

One of the major challenges in video captioning with object motion lies in the accurate estimation and tracking of object trajectories. Objects in videos often undergo occlusion, abrupt motion changes, or variations in scale and orientation, making it difficult to maintain consistent tracking. Additionally, differentiating between object motions (e.g., intentional movement versus passive displacement due to environmental forces) remains a complex task. Existing methods struggle to distinguish between these subtle variations, reducing the accuracy of motion-based captioning.

Another issue is the integration of motion information into captioning models. Many traditional [3][5] approaches rely on spatial feature extraction while incorporating only limited aspects of object motion. As a result, they fail to generate captions that capture the nature of object movements, such as acceleration, deceleration, collisions, or coordinated actions. Without explicitly modeling these motion types, generated captions often miss essential interaction details that could provide deeper insights into the scene.

Current object-oriented video captioning techniques leverage methods such as attention-based[16],[1],[3] graph-based, and estimation-based approaches. While these techniques have improved captioning quality, they primarily focus on scene-wide descriptions rather than object-specific movements. Graph-based models, for instance, extract object relationships but do not inherently account for motion characteristics. Attention mechanisms enhance localization but may not effectively[37]-[40] capture long-term object interactions or orientation shifts over time. Additionally, pose and shape estimation techniques provide useful spatial information but often fail to incorporate dynamic motion features, limiting their effectiveness in movement-focused video descriptions.

To address these challenges, our proposed approach focuses on enhancing motion-oriented feature extraction. By estimating object trajectories, identifying motion types, and analyzing movement paths, our method ensures that captions are not only spatially[23][26] accurate but also contextually rich in describing object interactions. Extracted motion features are mapped into a contextual framework, allowing for the generation of more meaningful and object-focused descriptions. This approach bridges the gap in capturing object-oriented motion information, ultimately improving the comprehensiveness and relevance of generated captions for various applications, including autonomous navigation, surveillance, and action recognition systems.

The method proposed in this article has demonstrated significant improvements in estimating object movements, their paths, and mapping to sample captions. This also provides insights into object interactions with other objects and the scene's background. The contributions made in this paper can be summarized as follows:

1. A mathematical model for estimating object path estimation across temporal aspects in the scene.
2. A specialized module for estimating object paths changes within the scene.
3. The creation of new object-oriented and movement oriented datasets containing video descriptions based on object interactions.
4. The conversion of features extracted by a CNN model into text in an object-oriented context.

This research article addresses critical aspects of object-oriented video captioning, enhancing our understanding of object interactions and motion within videos.

2. Related Work

The field of video captioning has evolved significantly, with modern techniques primarily centered around encoder-decoder architectures that aim to bridge the gap between visual content and linguistic representation. These models integrate various advanced methods, including graph-based learning, object-centric approaches, general transformer-based techniques, and estimation-driven methodologies. Typically, the encoding phase of these models employs Convolutional Neural Networks (CNNs) to capture static visual attributes and partial temporal dependencies, while the decoding stage is handled by Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks. This sequential pipeline ensures that visual and temporal cues are efficiently mapped into a contextual representation, which is then optimized using loss functions tailored to video captioning tasks [9]–[16].

Traditionally, video captioning has been largely dependent on spatial and temporal feature extraction techniques, which remain a dominant approach in research [4][7]. Some advanced methodologies have extended beyond these basics, delving into object-level fine-grained feature extraction and integrating attention mechanisms to refine caption quality [8][9][10]. Additionally, improvements have been made by incorporating localized feature importance processes using RNN-based architectures and other deep learning frameworks, enhancing the performance of general-purpose video captioning systems [18][19][21]. Attention mechanisms have also been widely applied at the decoding stage [22][23][24][26], allowing for improved object localization and better spatial feature representation when combined with RNN-based models. Further extensions of these techniques have explored multi-perspective object localization [28] and the shift towards generative captioning models instead of deterministic outputs [29]. These advancements, along with other research trends, have been extensively discussed in survey literature [30]. Another key area of innovation has been the simultaneous optimization of spatial and temporal feature extraction processes, enabling more efficient loss optimization and improved video captioning outcomes [32][33][35].

Graph-based methodologies have also played a vital role in video captioning by leveraging spatial and temporal relationships to construct structured feature graphs [1][6][11]. Once constructed, these graphs are further refined through learning algorithms that incorporate graph filters or CNN-based processing. Several concepts have emerged in this space, including scene graphs, multi-level learning hierarchies, and the integration of knowledge distillation techniques, where one graph model transfers learned insights to another [12][13][14]. The development of graph representations has been approached from multiple angles, such as extracting event-driven graphs in the temporal domain [17] and adopting teacher-student learning paradigms [20]. Additionally, hierarchical feature organization using non-standard convolutional filters has been investigated [25][27][30]. Although graph-based approaches demand substantial computational resources, they have demonstrated exceptional capability in generating object-focused video captions with enriched contextual understanding.

Object-centric feature extraction has gained traction as a means of focusing on objects across both spatial and temporal dimensions throughout an entire video sequence [2][3]. These methodologies often employ adversarial training techniques [5] or autoregressive modeling approaches [15] to enhance context vector mapping and optimize loss functions. Graph-based principles have also been adapted for object detection [31] and pose estimation [36], enabling models to infer object orientation and movement dynamics with greater precision. These

techniques, while computationally intensive, yield highly relevant outputs for applications where object-centric analysis is essential. Motion estimation techniques [41], along with unidirectional temporal difference methods [42], have been developed to accelerate object-based classification processes, thereby improving efficiency without compromising accuracy.

In addition to these high-level approaches, estimation-based techniques have been introduced to track object positions and movements more effectively. Some of these methods include local motion orientation estimation [37] and edge-based motion analysis [38], which focus on identifying general motion patterns rather than object-specific movements. These techniques are further leveraged to extract key features that facilitate the modeling of object-object and object-background interactions. Other motion estimation techniques, such as motion generation-based tracking [39] and grid-based shape estimation [40], have been developed to provide refined object representations through structured spatial positioning. Additionally, certain methodologies have been designed to enhance object shape estimation by iteratively refining object boundaries during motion sequences, reducing inaccuracies caused by blurring effects [41]. While these approaches deliver precise object tracking capabilities, they often demand significant computational resources, which can limit their scalability for real-time applications.

The effectiveness of these video captioning methodologies is commonly assessed using a variety of evaluation metrics. Some of the most widely used benchmarks include BLEU [43], METEOR [44], ROUGE [45], and CIDEr [46], each offering unique insights into the quality of generated captions. BLEU, for instance, is well-suited for evaluating short sentence structures where grammatical precision is less critical. In contrast, ROUGE, which measures n-gram occurrences, is particularly effective for assessing textual coherence in video captioning applications. METEOR employs a strict word-matching approach, making it valuable for evaluating linguistic accuracy, while CIDEr is often used to assess the consistency of frame-to-frame caption generation tasks. Importantly, a lower score on one of these metrics does not necessarily indicate an inferior model, as each metric evaluates different aspects of caption quality.

Despite these advancements, video captioning still faces several unresolved challenges, particularly concerning the integration of object motion and movement characteristics. Many existing models struggle with tracking objects accurately across long sequences, especially when objects undergo occlusion, scale variation, or sudden trajectory shifts. Additionally, distinguishing between different types of object motions—such as self-propelled movement versus externally influenced motion—remains a difficult task. Another challenge lies in the seamless integration of motion features with language models; most conventional methods emphasize spatial feature extraction while treating motion dynamics as secondary information, leading to captions that lack depth in describing object interactions. To address these issues, future research should focus on developing more robust motion-aware feature extraction techniques that can better capture object movements and their contextual implications within a video.

Finally video captioning has made remarkable progress through the adoption of encoder-decoder models, attention mechanisms, and graph-based learning techniques. Object-centric approaches and estimation-driven methods have further expanded the potential of these systems by improving motion estimation and object detection

capabilities. However, significant computational challenges and limitations in accurately modeling object interactions still persist. As research continues to advance, the development of more sophisticated motion-aware captioning frameworks will be key to enhancing the depth and contextual accuracy of generated descriptions. By addressing these challenges, future models will be better equipped to provide detailed and meaningful captions that accurately reflect both spatial structures and dynamic movements within a video.

3. Problem Description

The field of object-oriented video captioning focuses on tracking object motion and its interactions with the environment, including backgrounds and other objects. Current advancements in video captioning primarily depend on deep learning techniques to generate descriptive text based on spatial and temporal features. However, these models often prioritize producing general scene descriptions rather than object-specific narratives, limiting their effectiveness in applications requiring detailed object interactions. Additionally, the lack of dedicated datasets designed specifically for object-oriented captioning poses a significant challenge in developing more precise and meaningful descriptions. To overcome these limitations, researchers have explored movement estimation techniques for generating captions that focus on object behavior. Unlike learning-based models, these estimation methods do not require extensive labeled datasets. Instead, they rely on tracking object motion through optical flow, position changes, and trajectory estimation. While these methods provide an alternative to dataset-dependent deep learning approaches, they lack critical parameters such as accurate object estimation, precise motion paths, and interaction details. As a result, object movement estimation alone is often insufficient in capturing the full complexity of object dynamics within a video. Elements such as grouped activities, motion patterns, and object-object collisions remain difficult to model with purely estimation-based techniques.

A major challenge in object-oriented video captioning is understanding diverse motion patterns and object interactions. Objects in videos may exhibit periodic movements, irregular motion, or even collide with other objects, creating complex interaction scenarios. Despite their importance, these factors remain largely unexplored in mainstream video captioning models. Traditional deep learning-based methods, while effective in general video captioning, struggle with object-specific descriptions due to their reliance on broad spatial and temporal feature extraction. Moreover, training such models requires extensive datasets that explicitly contain object interactions, collisions, and movement patterns. The process of training these models is computationally expensive, requiring significant time and resources to adapt them to new domains and scenarios.

Given these limitations, there is a need for a hybrid approach that combines learning-based techniques with object estimation methods to improve object-specific video captioning. A hybrid model would leverage the strengths of both estimation-based and deep learning-based methods to provide more detailed and context-aware captions. By integrating pre-trained object detection models with motion estimation techniques, object movements and interactions can be analyzed without the necessity of specialized datasets. This approach would enable the generation of captions that better capture object trajectories, interactions, and motion types while reducing the dependency on large-scale annotated datasets. The proposed hybrid approach has significant potential across various real-world applications. For instance, in traffic monitoring, object-oriented captions could help detect and

describe anomalies, such as vehicles moving in the wrong direction or sudden collisions. Similarly, in drone surveillance, the system could generate captions that describe object activities with higher accuracy, enabling improved situational awareness. By addressing the current gaps in object motion estimation and captioning, this research contributes to advancing the field of video understanding and creating more practical and scalable captioning models.

4. Proposed Method:

This section describes the framework of our proposed approach, which is structured into four key modules, as illustrated in Figure 1:

1. Object Detection and Tracking: This module is responsible for identifying objects and tracking their positions throughout the video frames. It plays a crucial role in recognizing object types and spatial locations, serving as the basis for subsequent motion analysis.
2. Shift-Vector Computation: After detecting objects, their movement patterns are analyzed by computing shift vectors. This module evaluates whether objects follow similar or independent motion trajectories by comparing their speeds and movement directions.
3. Path and Position Estimation: Based on the computed shift vectors, this module predicts the future positions and motion paths of objects. It facilitates understanding of object interactions and different motion types within the scene.
4. Caption Generation via Transformer Model: Extracted motion-related features are processed using a CNN-based encoder integrated into a transformer model. This final module generates object-centric captions describing the observed movements and interactions.

To effectively extract motion parameters essential for object-oriented video captioning, the workflow is divided into these four systematic modules. The first module ensures precise object localization across frames. The second module computes motion shifts, categorizing movements accordingly. The third module estimates object motion types by considering predefined parameters. Finally, the fourth module converts these motion patterns into meaningful text-based descriptions.

The object detection step in the first module is performed using two advanced models: FastTrackRcnn [47] and Detect Everything [48]. These models efficiently detect and localize objects, providing accurate data for further motion analysis in subsequent modules.

For second module let's consider the video file with certain frame rate fs. The video signals are converted into grayscale and its frames are defined as two-dimensional matrix $M[a, b, c]$. The pixels inside the matrix M at a point m has intensities which can be defined as $m = (m1, m2, m3)$.

Furthermore, to the above basic input parameters we have extracted objects **O** denoted by small matrix which hold the object $So(a,b,c)\ where\ o2\ \leq o2\ \leq O2$. This extracted object feature vector $S(o)$ is the one in which we are interested, this will enable us to calculate its displacement of an object in the scene. Let $\delta i[nm] = (\delta i, 1[n], \delta i, 2[n])$ is the displacement vector and its array, which contains the displacement of the i[th] object in scene (0 to nth frame). Once the displacement vector is modelled, calculation for the shift of the object in 2D space is to be modelled. So, as the displacement vector is calculated of the i[th] object, the shift for the i[th] object is S_i [m- $\delta i[n]$] for n frames. Where m is the intensity matrix for object features as described above. Once the intensity shift for the i[th] object is calculated, the intensity shift for the full video is to be modelled as follows.

$$M[a,b,c] = \sum_{i=0}^{O} Si[m - \delta i[n], n\ ] + B[m\text{-}n] + N[m] \quad (1)$$

Here, the frame matrix M[a,b,c] is equivalent to the sum of all the shifts vectors and the background matrix B[m] with some noise components N[m] arised from estimation erros. This noise is assumed to be Gaussian noise for the experimental only purposes , which are demonstrated in the below sections.

To simplify the above model above, basic limiting assumptions are considered for the speed of the objects array and also the periodicity of the object arrayed motions. Let's say we have similar speeds of the objects in linear domain then the displacement vector can be reduced to $\delta i[n] = vn$. The speed v can be in 1D or 2D as v = (v1, v2). Now (1) is simplified as (2) below.

$$M[a,b,c] = \sum_{i=0}^{O} Si[m - vn, n\ ] + B[m\text{-}n] + N[m] \quad (2)$$

Another level of simplification is the use of the constant background B[m] as a constant value b. So, the (2) will be written as (3)

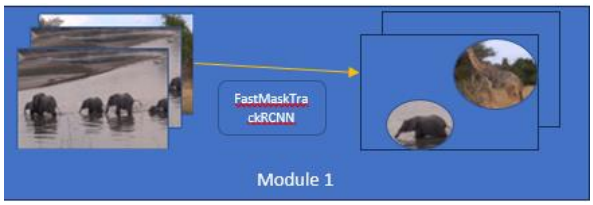
$$M[a,b,c] = \sum_{i=0}^{O} Si[m - vn, n*m\ ] + N[m\text{-}n] + b \quad (3)$$

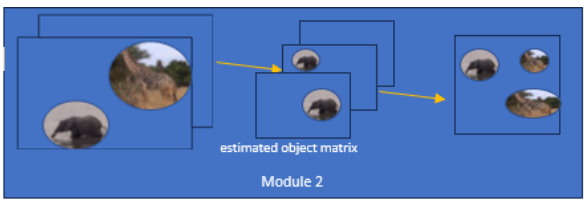
Conversion of the above equation will be better in time domain for calculations of background interaction and noise addition.

$$Mx[k,n] = \sum_{i=0}^{O} Si\ [k,n]e^{-ivnt} + W[k,n] \quad (4)$$

Now, that estimators are calculated. For the sake of simple calculations for this article, we have the case of two object synthetic video and other video with background removed. In particular, the test of the motion for its periodicity and collision will be tested and experimented in the experimentation section.

Fourth module takes the motion vectors from the third module and the objects features from the first module and maps to train for producing captions. This model requires an encoder decoder approach for mapping the context vector extracted using cnn-rnn based network from the matrix of motion features. This mapping requires cross entropy loss calculation and gradient decent method for reducing the errors.





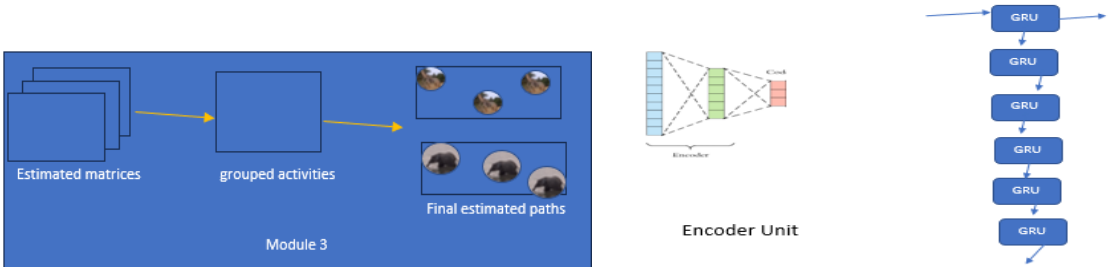


Figure 1: Overall Structure of proposed approach

Figure 1 describes the overall system design in which four modules are designed for dedicated tasks. As explained in the section these modules are implemented separately.

5. Experimentations

The experiments para introduces the working setup and the dataset curated for the this article on video captioning. new dataset with changes of base truth sentences , updated truth is presented in Figure2. The results are reported in tabular form for quantitative comparison. Other results are presented in Figures 3,4. Finally, a detailed ablation study with variations in the model and its types is presented in the next section.

The comparison in table no.s 2,3,4 is compared on 3 main metrics as the score of metrics used with an estimated number of objects and conditions of background. The experiments required a training system with a configuration, which suits the input shape and output shape characteristics of the article. The system used for training and estimation is NVIDIA A100 Tensor Core GPU. This system consists of 10 tensor core GPUs which have 10 parallel docker's instances to run. The system runs with 2TB of memory and 30TB of SSD storage. This paper utilizes a docker instance, which runs on one of the GPU cores and has 100GB of dedicated memory. The docker instance runs on a PyTorch v1.9 with Python 3.6. The Jupyter notebooks run the ver. 7 as it is more flexible than Jupyter 4 for storing the outputs into a separate file.

The dataset requirements for this article are different and require to be prepared manually just like [3]. The dataset requires video-to-text mapping in a way that describes the object and its motions along with other aspects of object oriented-ness. The article scoped itself to 55 videos for estimation & training and 25 videos for testing and score calculation purposes. To ensure the unbiases of activities like linear motion and collisions, it is chosen with utmost care with multiple activity-based videos and subsequent object-related captions. These videos are taken from UCF101 and MSVD databases. These videos are tested for keeping the bucketed context of the videos as per P.kaushik [2] in its video description part. This allows the text context to not go on a very wide range, it will be limited to limited activities. Some other details of the dataset are listed in the below table.

Table 1: Comparison of datasets with standard datasets for video understanding training.

Dataset	Task	Type	Length in sec	objects	Length of sentence	Verb per sentence	Verb ratio	Adjective per sentence
MSR-VTT	Video captioning	Clip-sent	20	3	9.28	1.37	14.80	0.66
MSVD	Video captioning	Clip-sent	10	3	8.67	1.33	19.60	0.25
AcivityNet	Dense Video captioning	Clip-sent	180	3	13.48	1.41	10.40	0.67

F.Zhu[3]	Object-oriented Video captioning	Object-sent	73	3	16.56	2.02	21.00	1.97
Our Dataset	Object-oriented Video captioning	Object-sent	56	4	20.2	2.11	22.00	1.33

Object oriented caption- Some elephant are walk over water,
GT- Elephants are passing through river

Object oriented caption- two giraffe are passing in trees
GT - some giraffes are walking in the garden

Figure 2: some sample object-oriented video frames and cations

Various evaluation metrics are used to assess the quality of generated video captions, including ROUGE [45] and BLEU [43], among others. This paper employs multiple evaluation techniques to ensure a comprehensive assessment. Specifically, Word Mover Distance (WMD) is utilized due to its effectiveness in summarization tasks. BLEU is applied to measure the numerical word closeness between detected objects and their motions, while CIDEr [46] is used to evaluate cosine distance-based scores, offering insights into caption accuracy in a contextual manner.

In addition to quantitative evaluations, qualitative comparisons are also presented, allowing for a visual examination of the differences between generated captions. Moreover, an ablation study is conducted in the following section to analyze the impact of various model components by selectively adding or removing certain elements and observing their effects on the results.

To ensure a comparative analysis with state-of-the-art methods, this paper benchmarks its results against the works of F. Zhu [3], Li L [4], S. Venugopal [19], J. Kotera [41], and Aiden SJ [35]. Some previous studies, such as [41], have employed METEOR and ROUGE-based metrics, treating captioning as a summarization task rather than frame-wise captioning. This paper, however, takes an intermediate approach by incorporating motion-based temporal features, making it distinct from purely summarization-based or frame-by-frame methods.

For a structured comparison, this paper evaluates results using BLEU [43], ROUGE [45], and CIDEr [46], ensuring alignment with state-of-the-art methodologies. A detailed ROUGE-based comparison of the generated captions is presented in Table 2.

Table 2: Comparison of ROUGE based metric results

Model	**ROUGE-L**	**Avg. Number of objects**	**Background changes**
F.Zhu [3]	47	~3	Yes

Li L [4]	39	3	Yes
S. vennugopan [19]	35	NA	NA
J Kotera [41]	40	NA	Yes
Aiden SJ [35]	42	~3	Yes
Proposed method	**43.1**	**~5**	**3 times**

Table 2 presents the results of the proposed method for generating video captions using **object motion estimation analysis**. Unlike conventional state-of-the-art methods, which focus on **static object-oriented captioning**, this approach explicitly incorporates **motion aspects** and **background interactions** to enhance caption quality. Existing methods primarily rely on encoder-decoder models, where background interactions are extracted implicitly rather than being a core feature. In contrast, the proposed approach ensures **dynamic and context-aware captions**. The table below provides a **BLEU-based comparison** with selected methods, along with a **ROUGE-based evaluation**, highlighting the improvements in motion-aware video captioning.

Table 3: Comparison of BLEU based metric results

Model	BLEU@1	Avg. Number of objects	Background changes
F.Zhu [3]	52	~3	Yes
Li L [4]	39.3	3	Twice
S. vennugopan [19]	36	NA	NA
J Kotera [41]	50	NA	Yes
Aiden SJ [35]	39	~3	Yes
Proposed method	**50.1**	**~5**	**3 times**

Below are the results for the CIDEr based comparison with methods chosen above and compared with ROUGE based metrics.

Table 4: Comparison of CIDEr based metric results

Model	CIDEr-D	Avg. Number of objects	Background changes
F.Zhu [3]	47	~3	Yes
Li L [4]	39	3	Twice
S. vennugopan [19]	35	NA	NA
J Kotera [41]	40	NA	Yes
Aiden SJ [35]	42	~3	Yes
Proposed method	**44.1**	**~5**	**3 times**

6. Ablation Study

The ablation study is conducted to evaluate the significance of each component within the proposed method by analyzing various sub-modules and comparing them with alternative approaches. Such studies help identify key contributing factors that lead to performance improvements.

The results of modifying the model's hyperparameters are presented in Tables 5 and 6, highlighting comparisons based on estimated object count, training frames, loss functions, and computational FLOPs. The study focuses on Modules 2 and 3, which handle object motion estimation, while Module 4, a transformer-based module, uses Cross-Entropy Loss (CEL) and Focal Loss (FL) for training. The tested models and results are detailed below.

a. Base-Model-1: The baseline model trained with both kinds of loss functions CEL and FL. This includes 1000 frames.
b. Base-Model-2: This baseline model is trained for estimated objects that are categorized with below 5 below 8 etc.
c. Base-Model-3: This baseline model is trained for estimated objects above 6 etc.
d. Base-Model-4: This baseline model is trained for estimated objects above 6 etc.
e. Base-Model-5: This baseline model is trained for estimated objects above 8 etc.
f. Base-Model-6: This baseline model is trained for estimated objects above 8 etc.

The Base model is categorized as above with various parameters like the number of objects estimated. Frames number and loss functions. The estimated objects are as mentioned and the number of objects present in the video may differ. We have taken the estimated objects as parameters, not the number of objects present in the video.

Table 5: Comparison of various models with two different loss functions and number of objects.

S.No	Model	BLEU@4	CIDEr	ROUGUE-L	N as number of Frames	O as number of objects	Loss	FLOP(e^{10})
1	Base-line-1	50.1	45.3	45.1	1000	< 5	CEL	0.4
2	Base-line-2	49.1	45.3	46.4	1000	< 6	FL	0.5
3	Base-line-3	49.1	45.5	46.1	1000	> 6	CEL	0.7
4	Base-line-4	49.0	45.3	42.1	1000	> 6	FL	0.7
5	Base-line-5	50.3	45.5	38.5	1000	>8	CEL	0.8
6	Base-line-6	50.3	45.5	38.7	1000	>8	FL	1.01

The above table 5 results are when the background of the videos is not changing much, so the features extracted from the background and object are very less. Now lets see the results of the `test cases where the background is also changing within the video. These kinds of videos are mostly the videos which shot for moving objects like cars, etc not from the static CCTV based.

Table 6: Comparison of various models with changing background and reduced frames.

S.No	Model	BLEU@4	CIDEr	ROUGUE-L	N as number of Frames	O as number of objects	Loss	FLOP(e^{10})
1	Base-1bg	42.1	33.1	40.1	500	< 5	CEL	1.1
2	Base-2bg	40.1	33.3	40.4	500	< 6	FL	1.2
3	Base-3bg	40.1	33.3	34.1	500	> 6	CEL	1.1
4	Base-4bg	40.0	36.3	34.1	500	> 6	FL	1.2
5	Base-5bg	42.7	42.9	36.1	500	>8	CEL	1.5
6	Base-6bg	42.8	43.9	36.5	500	>8	FL	1.0

The table 6 used the test videos with changing background which also have some objects and are also estimated and shows their effects in computation intensiveness specially for the last two models. Likewise Figure 3, 4 shows the results in qualitative samples.

paths of objects

Figure 3a.
Object oriented Ground Truth: Basic shapes moving forward
GT from Proposed method: Two basic shapes are moved in forward.

paths of objects

Figure 3b.
Object oriented Ground Truth: butterflies moves
GT from Proposed method: butterflies are moves on base floor

Figure 3: Qualitative comparison results various parameters and models without changing background

paths of objects

General Ground Truth: animals walks in forest
Object oriented Ground Truth: two giraffes walking in woods
GT from Proposed method: giraffes walks around trees.

Figure 4: Qualitative comparison results various parameters and models with changing backgrounds

Figure 3a illustrates sample frames from a synthetic video dataset featuring a fixed background and two moving objects. The proposed method successfully extracted and plotted the object paths as intermediate results, leading to the final caption generation. The ground truth describes the scene as "basic shapes moving," whereas the generated caption provides additional insight: "Two basic shapes are moving forward," incorporating motion direction.

Figure 3b presents a real-world video with a stable background. Like Figure 3a, the object paths were extracted and analyzed using the proposed method. The generated caption introduced additional context by stating, "butterflies are moving on the ground," whereas the ground truth only noted butterfly movement without specifying the surface. This demonstrates the method's ability to enhance video understanding by capturing finer details that may be missing from the original annotations.

Figure 4 introduces a more complex scenario by using a video with a changing background. The proposed method is compared with object-oriented methods to evaluate its effectiveness. In this case, both approaches detected the background and incorporated it into the generated captions. However, a key difference is that object-based methods do not consider object paths, whereas the proposed method integrates path estimation for a richer scene understanding. By incorporating motion paths, the proposed approach provides a more detailed and comprehensive description of object interactions within the scene.

7. Conclusion

This study introduced a novel approach for object-oriented video captioning by leveraging estimation of object shift vectors. The proposed method effectively improves object-focused caption generation, demonstrating its advantages over traditional techniques. A unique dataset comprising 100 videos was curated specifically for this research, offering an object-activity-sentence tuple to enhance object-based video understanding. Various model variations were tested, and their impact on both computational efficiency and caption accuracy was thoroughly analyzed. While previous research has explored object-oriented analysis and estimation-based techniques independently, this study is the first to integrate both approaches for video caption generation. The estimation-based methodology provided a deeper analysis of object interactions within the spatial and temporal domains. Experimental results validated the effectiveness of this method, highlighting its capability to deliver scene descriptions from an object-centric perspective. Object-oriented video captioning is gaining increasing attention due to its applicability in industrial automation, such as detecting inappropriate content, monitoring machine malfunctions, collision detection, and identifying traffic violations like wrong-way driving. The combination of estimation techniques and advanced transformer-based models is proving to be highly effective in these domains. However, a major challenge in this field remains the limited availability of relevant datasets, prompting researchers to explore unsupervised learning techniques as an alternative. With growing industrial demand for automated video analysis and object-based understanding, the future holds great potential for estimation-driven

captioning methods. Continued advancements in dataset development, unsupervised learning, and transformer-based architectures will further enhance the accuracy and applicability of object-oriented video captioning in real-world scenarios.

References

[1] Lin, K., Li, L., Lin, C.C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y. and Wang, L., 2022. 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).

[2] P. Kaushik, K. V. Kumar and P. Biswas, "Context Bucketed Text Responses using Generative Adversarial Neural Network in Android Application with Tens or Flow-Lite Framework," *2022 8th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2022, pp. 324-328, doi: 10.1109/ICSC56524.2022.10009634.

[3] F. Zhu, J. -N. Hwang, Z. Ma, G. Chen and J. Guo, "Understanding Objects in Video: Object-Oriented Video Captioning via Structured Trajectory and Adversarial Learning," in *IEEE Access*, vol. 8, pp. 169146-169159, 2020, doi: 10.1109/ACCESS.2020.3021857.

[4] Li, L. et al. (2022) 'Adaptive spatial location with balanced loss for video captioning', IEEE Transactions on Circuits and Systems for Video Technology, 32(1), pp. 17–30. doi:10.1109/tcsvt.2020.3045735.

[5] X. Hua, X. Wang, T. Rui, F. Shao and D. Wang, "Adversarial Reinforcement Learning With Object-Scene Relational Graph for Video Captioning," in *IEEE Transactions on Image Processing*, vol. 31, pp. 2004-2016, 2022, doi: 10.1109/TIP.2022.3148868.

[6] Hendria, W.F. et al. (2023) 'Action knowledge for video captioning with graph neural networks', Journal of King Saud University - Computer and Information Sciences, 35(4), pp. 50–62. doi:10.1016/j.jksuci.2023.03.006.

[7] Moniruzzaman, M. et al. (2022) 'Human action recognition by discriminative feature pooling and video segment Attention Model', IEEE Transactions on Multimedia, 24, pp. 689–701. doi:10.1109/tmm.2021.3058050.

[8] Yu, Q., Song, J., Song, YZ. et al. Fine-Grained Instance-Level Sketch-Based Image Retrieval. Int J Comput Vis 129, 484–500 (2021). https://doi.org/10.1007/s11263-020-01382-3

[9]] C. Yan et al., "STAT: Spatial-Temporal Attention Mechanism for Video Captioning," in IEEE Transactions on Multimedia, vol. 22, no. 1, pp. 229-241, Jan. 2020, doi: 10.1109/TMM.2019.2924576.

[10] Prudviraj, J. et al. (2022) 'AAP-MIT: Attentive Atrous Pyramid Network and memory incorporated transformer for multisentence video description', IEEE Transactions on Image Processing, 31, pp. 5559–5569. doi:10.1109/tip.2022.3195643.

[11] Z. Zhang, D. Xu, W. Ouyang and L. Zhou, "Dense Video Captioning Using Graph-Based Sentence Summarization," in *IEEE Transactions on Multimedia*, vol. 23, pp. 1799-1810, 2021, doi: 10.1109/TMM.2020.3003592.

[12] Yan, Y. et al. (2022) 'Fine-grained video captioning via graph-based multi-granularity interaction learning', IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(2), pp. 666–683. doi:10.1109/tpami.2019.2946823.

[13] Lu, X. and Gao, Y. (2022) 'Guide and interact: Scene-graph based generation and control of video captions', Multimedia Systems, 29(2), pp. 797–809. doi:10.1007/s00530-022-01012-7.

[14] B. Pan, et al., "Spatio-Temporal Graph for Video Captioning With Knowledge Distillation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020 pp. 10867-10876. doi: 10.1109/CVPR42600.2020.01088

[15] Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. 2021. O2NA: An Object-Oriented Non-Autoregressive Approach for Controllable Video Captioning. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 281–292, Online. Association for Computational Linguistics

[16] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," IEEE Trans. Multimedia, vol. 21, no. 8, pp. 2117–2130, Aug. 2019.

[17] Wang, T., Zheng, H., Yu, M., Tian, Q., & Hu, H. (2021, May). Event-Centric Hierarchical Representation for Dense Video Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(5), 1890–1900. https://doi.org/10.1109/tcsvt.2020.3014606

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[19] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell and K. Saenko, "Sequence to Sequence -- Video to Text," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4534-4542, doi: 10.1109/ICCV.2015.515.

[20] Z. Zhang et al., "Object Relational Graph With Teacher-Recommended Learning for Video Captioning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 13275-13285, doi: 10.1109/CVPR42600.2020.01329.

[21] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 1 April 2017, doi: 10.1109/TPAMI.2016.2599174.

[22] J. Xu, T. Mei, T. Yao and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 5288-5296, doi: 10.1109/CVPR.2016.571.
[23] L. Gao, Z. Guo, H. Zhang, X. Xu and H. T. Shen, "Video Captioning With Attention-Based LSTM and Semantic Consistency," in IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2045-2055, Sept. 2017, doi: 10.1109/TMM.2017.2729019.
[24] R. Krishna, K. Hata, F. Ren, L. Fei-Fei and J. C. Niebles, "Dense-Captioning Events in Videos," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 706-715, doi: 10.1109/ICCV.2017.83.
[25] H. Yu, J. Wang, Z. Huang, Y. Yang and W. Xu, "Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4584-4593, doi: 10.1109/CVPR.2016.496.
[26] K. Cho, A. Courville and Y. Bengio, "Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks," in IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 1875-1886, Nov. 2015, doi: 10.1109/TMM.2015.2477044.
[27] P. Pan, Z. Xu, Y. Yang, F. Wu and Y. Zhuang, "Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1029-1038, doi: 10.1109/CVPR.2016.117.
[28] J. Yu, J. Li, Z. Yu and Q. Huang, "Multimodal Transformer With Multi-View Visual Representation for Image Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp. 4467-4480, Dec. 2020, doi: 10.1109/TCSVT.2019.2947482.
[29] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic and H. T. Shen, "From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 10, pp. 3047-3058, Oct. 2019, doi: 10.1109/TNNLS.2018.2851077.
[30] Islam, S., Dash, A., Seum, A. *et al.* Exploring Video Captioning Techniques: A Comprehensive Survey on Deep Learning Methods. *SN COMPUT. SCI.* **2**, 120 (2021). https://doi.org/10.1007/s42979-021-00487-x
[31] Li, J., Tan, G., Ke, X. *et al.* Object detection based on knowledge graph network. *Appl Intell* **53**, 15045–15066 (2023). https://doi.org/10.1007/s10489-022-04116-9
[32] Liu, Zy., Liu, Jw. Hypergraph attentional convolutional neural network for salient object detection. *Vis Comput* **39**, 2881–2907 (2023). https://doi.org/10.1007/s00371-022-02499-x
[33] Chen, S., Zhong, X., Li, L. *et al.* Adaptively Converting Auxiliary Attributes and Textual Embedding for Video Captioning Based on BiLSTM. *Neural Process Lett* **52**, 2353–2369 (2020). https://doi.org/10.1007/s11063-020-10352-2
[34] Su, Y., Li, Y., Xu, N. *et al.* Hierarchical Deep Neural Network for Image Captioning. *Neural Process Lett* **52**, 1057–1067 (2020). https://doi.org/10.1007/s11063-019-09997-5
[35] Jo, Yongrae, Seongyun Lee, Aiden SJ Lee, Hyunji Lee, Hanseok Oh, and Minjoon Seo. "Zero-Shot Dense Video Captioning by Jointly Optimizing Text and Moment." *arXiv preprint arXiv:2307.02682* (2023).
[36] T. Yasunaga, T. Oda, N. Saito, A. Hirata, K. Toyoshima and K. Katayama, "Object Detection and Pose Estimation Approaches for Soldering Danger Detection," *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, Kyoto, Japan, 2021, pp. 697-698, doi: 10.1109/GCCE53005.2021.9621849.
[37] R. Kalboussi, A. Azaza, M. Abdellaoui and A. Douik, "Detecting Video Saliency via Local Motion Estimation," *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, Hammamet, Tunisia, 2017, pp. 738-744, doi: 10.1109/AICCSA.2017.93.
[38] M. Asikuzzaman, A. Ahmmed, M. R. Pickering and T. Sikora, "Edge Oriented Hierarchical Motion Estimation For Video Coding," *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020, pp. 1221-1225, doi: 10.1109/ICIP40778.2020.9190852.
[39] T. Kitamura, X. Sun, Y. Saito, H. Asai, T. Nozaki and K. Ohnishi, "Motion Generation Based on Physical Property Estimation in Motion Copy System," *2022 IEEE 17th International Conference on Advanced Motion Control (AMC)*, Padova, Italy, 2022, pp. 62-67, doi: 10.1109/AMC51637.2022.9729324.
[40] S. Steyer, C. Lenk, D. Kellner, G. Tanzmeister and D. Wollherr, "Grid-Based Object Tracking With Nonlinear Dynamic State and Shape Estimation," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 2874-2893, July 2020, doi: 10.1109/TITS.2019.2921248.
[41] J. Kotera and F. Šroubek, "Motion Estimation and Deblurring of Fast Moving Objects," *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 2860-2864, doi: 10.1109/ICIP.2018.8451661.
[42] P. Kaushik and V. Saxena, "Fast Video Classification based on unidirectional temporal differences based dynamic spatial selection with custom loss function and new class suggestion," *2023 International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2023, pp. 419-423, doi: 10.1109/ICDT57929.2023.10150644.
[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation,in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2001, pp. 311318.
[44] S.BanerjeeandA.Lavie,Meteor:AnautomaticmetricforMTevaluation with improved correlation with human judgments,in Proc. ACL Work shop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization, 2005, pp. 6572.
[45] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries,in Proc. Text Summarization Branches Out, 2004, pp. 7481.

[46] R. Vedantam, C. L. Zitnick, and D. Parikh, CIDEr: Consensus-based image description evaluation,in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 45664575.
[47] M. Maity, S. Banerjee and S. Sinha Chaudhuri, "Faster R-CNN and YOLO based Vehicle detection: A Survey," *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2021, pp. 1442-1447, doi: 10.1109/ICCMC51019.2021.9418274.
[48] M. S. S. Reddy, P. R. Khatravath, N. K. Surineni and K. R. Mulinti, "Object Detection and Action Recognition using Computer Vision," *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2023, pp. 874-879, doi: 10.1109/ICSCSS57650.2023.10169620.