A comprehensive review of CNN and transformer based visual feature extraction for automatic video captioning

Hemel Sharker Akash [a,1], Joseph Emerson Raja [a,2,*], Md. Jakir Hossein [a,3]

[a] Faculty of Engineering & Technology (FET), Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
[b] Second affiliation, Address, City and Postcode, Country (9pt)
[1] hemelakash472@gmail.com; [2] emerson.raja@mmu.edu.my; [3] jakir.hossen@mmu.edu.my
* corresponding author

ARTICLE INFO

Article history
Received November 6, 2025
Revised April 8, 2026
Accepted April 9, 2026
Available online August 31, 2026

Keywords
Video Captioning
Video feature extraction
Vision-Language Models
CNN
Attention mechanism

ABSTRACT

Automatic Video Captions are increasingly important to use in applications in the accessibility, search, education, and content moderation areas. Generating accurate and context-aware captions is difficult for lengthy and complex videos. Recent surveys focus more on language models than on the vision side. This means that the contribution of visual feature extraction towards the performance of the model is not studied properly. Here we attempt to fill this gap by conducting reviews of video captioning literature from the vision viewpoint, comparing CNN-based and transformer-based feature extraction, and detecting trends, pros, and cons. Approximately 100 papers published between 2017 and 2025 were reviewed and analyzed based on the base vision, dataset, and evaluation metrics. The review classified the selected papers into CNN-based pipelines and transformer-based techniques. Moreover, it compared their rank measured by application popularity, dataset usage, and performance scores across popular datasets/benchmarks. The review shows that CNNs were dominated before 2021, but transformer-based approaches now lead due to their ability to capture long-range temporal dependencies and multimodal interactions. However, CNNs continue to perform well for short videos. As for Dataset usage, patterns show that ActivityNet Captions and YouCook2 are suited for long-form reasoning because of timely manner captioning, while MSR-VTT and MSVD are effective for short clips (one sentence caption). In terms of Special semantic quality metrics, CIDEr and METEOR are more semantically appropriate compared to n-gram quality metrics like BLEU. This review emphasizes transformer-based architecture, particularly multimodal models, which will be the future of video captioning. However, challenges remain in handling lengthy videos, adapting across domains and languages, and ensuring efficient deployment (small, optimized model). Addressing these gaps will add value by making video captioning systems more reliable, interpretable, and broadly usable in real-world contexts.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

Automatic video captioning (AVC), or video (visual actions or activities) to text generation. It is one of the methods of teaching computers to describe moving images in natural language. Image captioning, when models describe the event only out of one frame, because of that, it occasionally misinterprets the surrounding area or even. But AVC is not such. It breaks down each moment frame by frame, and then describes the moment or event incredibly precisely of the video, since this time model can see more surrounding areas and establish more connection between events. The uses are not hard to imagine.

Hours of video could be scanned by search engines within a few seconds, moderators would do their job more efficiently, and teachers could make the summaries of lectures without making an effort.

In comparison to the modern day, the primitive machines were cumbersome. They used manually chosen features or inflexible templates, and this resulted in almost monotonous and monosyllabic descriptions. The retrieval techniques were slightly better by copying the caption in a similar-looking clip, but when something was changed (either a change of color or a change of shape), the system did not find any match, and the system failed. It changed with the entrance of deep learning. CNNs [1] were selected as the default to extract spatial and short-term motion features (previously CNN was the most effective at extracting features in images), typically combined with RNNs such as LSTMs to chain those features into text. These pipelines served well with short clips, but when videos became longer or the action more complicated, the models lost track of what had already transpired. In essence, it forgot the past action that occurred many years ago.

Transformers [2] changed the landscape again. They had already proven themselves in natural language processing, and soon vision researchers adapted them into models like Vision Transformer [3], TimeSformer [4] and Video Swin. Unlike RNNs, Transformers could connect events across distant frames without grinding step by step through every image. On benchmarks, they began overtaking the older CNN–RNN pipelines. More recently, multimodal large language models such as BLIP-2 [5], Flamingo [6], Video-LLaVA [7] have pushed the idea further by trying to combine visual and linguistic reasoning inside a single Transformer-based framework. They are still new and not yet the standard, but the field seems to be moving in their direction. Even so, most surveys [8], [9], [10], [11], [12], [13], [14], [15] of video captioning still lean heavily toward the language side: attention mechanisms, datasets, or decoder design. The visual part of the pipeline, the step that decides what the model "sees" before it ever speaks, often gets little more than a paragraph. And while MLLMs are sometimes mentioned, they rarely receive a proper slot in existing taxonomies or side-by-side comparisons.

This review tries to shift the focus back toward vision. We look at CNN-based and Transformer-based feature extraction, track how their roles have changed between 2017 and 2025, and examine the datasets and evaluation metrics most researchers rely on. Our survey covers about 100 peer-reviewed papers. The goal is not to deliver the final word, but to sketch where vision-focused strategies have brought the field so far, where they continue to fall short, and which directions, though uncertain, seem most promising for the years ahead.

The earliest systems for video captioning were built on two basic tricks: templates and retrieval [16]. This approach tried to find exact moments or kind of similar moments and describe them as fixed sentences. A typical output might be something like "A person is walking in a park." It was easy to implement, but the captions quickly started to sound repetitive and flat. Retrieval-based methods tried a different angle. Instead of filling in templates, retrieval methods tried to play matchmaker by scouring a database for a video that looked roughly similar and then reusing its caption. This at least produced grammatical sentences, but the moment a video showed something unusual or new, the system had no clue what to do. Both of these approaches also leaned heavily on handcrafted features like SIFT [17], HOG [18], or dense trajectories. Looking back, those features were never really built to capture the messy, overlapping events you get in real-world video. At least it produced grammatical sentences, but the moment a video showed something unusual or new, the system had no clue what to do. Both of these approaches also leaned heavily on selected features like SIFT, HOG, or dense trajectories. Looking back, those features were never really built to capture the messy, overlapping events you get in real-world video.

DL (deep learning) redefined the entire process so it does not depend on pre-selected features, but is trained to find the features automatically. The generic recipe was a CNN for vision, an RNN for text. Using 2D CNN-RNN models, one can extract image features with pre-trained ImageNet [19] backbones like VGG [20] or ResNet [21], average or pool them across time, and input them into an LSTM or GRU decoder. This was further expanded with three-dimensional CNNs, eg., C3D [22] and I3D [23],

which can extract features in a temporal manner. Hybrid models attempted to integrate the fine spatial detail of 2D CNNs with the motion modeling of 3D CNNs. Other researchers added audio encoders to record background noise or voice. The CNN-based models were better than the previous handcrafted era. Nonetheless, these systems continued to prefer to work on fixed-length clips and frequently lost long-range dependencies, such as attempting to recall the beginning of a movie having only watched the middle.

Attention mechanism [2], which allow models to determine where and when to look. Temporal attention [24] enables the decoder to pay attention to certain frames or moments; spatial attention to particular regions or objects. In this potent model, scientists have started to add additional modalities: the audio characteristics of models such as VGGish, or object localisations of Faster R-CNN [25] and DETR. This attention mechanism helps to build a more accurate language model. It is enhanced by future researchers to extract visual features. This multimodal combination frequently enriches captions, particularly in dense video captioning, when you are describing several events in one, unedited sequence. Attention still, however, was not a magic bullet to the problem of extracting information over very long and complex videos. The assisting hand of the self-attention mechanism transformed everything. Recurrence was substituted with self-attention, which enabled the models to correlate events occurring in the beginning and the end of a sequence. Spatial encoding with Vision Transformers (ViT) [3] and hierarchies like Swin Transformer [26] became popular, whereas temporal Transformers like TimeSformer [26] became popular for spatial encoding, while temporal Transformers such as TimeSformer [4] and Video Swin directly used spatiotemporal inputs. Soon, fully Transformer-based encoder–decoder systems came in the field, blending video and text into a single modeling space. After 2021, the shift became noticeable. Nowadays, most researchers are using transformer-based encoders rather than CNN-based visual feature extraction, supported by the growing availability of large-scale pretraining datasets.

Most recently, Multimodal Large Language Models (MLLMs) such as BLIP-2 [5], Video-LLaVA [7], Video-LLaMA [27], Ma-LMM [28], and others have pushed the idea further, aiming for unified vision-language reasoning. These models used transformer-based vision encoders. Training these models needed massive image-text or video-text datasets. Where decoder models that can handle richer text generation and even follow instructions. They've shown surprisingly strong zero-shot and few-shot results, meaning they can generate captions without much (or any) fine-tuning. We classify MLLMs under the transformer umbrella, as their visual backbones almost always rely on transformer encoders. However, this model is not well known because it was introduced recently. Their rapid growth indicates that they are likely to become a more visible presence in future research.

Several survey papers have explored AVC and related vision-language tasks. Most of these works describe or review the visual as well as the language side, or focus only on the language side, such as decoder designs, attention mechanisms, and evaluation metrics. While these aspects are important, the visual feature extraction stage tends to receive less attention, even though video feature extraction or encoding the video is one of the most important parts. In many surveys, visual processing is often treated as an initial step rather than a central component of the system. Because of this, there is limited in-depth discussion on how different visual encoders, including CNNs and transformer-based models, affect performance across various datasets and applications. At the same time, the recently developed MLLMs have not yet been fully discussed in most survey frameworks. Although some studies mention MLLMs, they are usually discussed separately and are not compared directly with earlier approaches in a consistent way.

In order to overcome these limitations, we concentrated on the video feature extraction part. Which models are most recently used? What kind of video encoder will be suitable for what purpose? Moreover, we unite CNN-based and transformer-based models. The MLLMs framework on which we are based is also based on transformers, thus simplifying the comparison of the various types of approaches. Table 1 clearly indicates how this work is different from previous surveys. It offers a systematic comparison along some of its requirements, such as the research focus, design of taxonomy, integration of MLLMs, and the degree of analysis of visual features. This survey offers a narrower and more systematic vision, as it

focuses more on visual processing, has a better taxonomy, and offers a more coherent analysis of existing solutions. This helps address several gaps that remain in the current literature.

Table 1. Comparison of Existing Video Captioning Surveys.

Survey	Year	Focus	Taxonomy	MLLMs Included	Vision Analysis Depth	Key Limitation
Islam et al.[8]	2021	Video captioning (deep learning)	CNN–RNN-based architectures	No	Low	Focuses mainly on architecture, limited analysis of visual feature extraction
Akhare and Shinde [9]	2022	Query-focused video summarization	Task-based taxonomy	No	Low	Focus on summarization rather than caption generation
Liu et al. [10]	2023	Video moment localization	Task-specific taxonomy	No	Low	Not focused on captioning; limited relevance to feature extraction for captions
Abdar et al.[11]	2024	Deep learning for video captioning	General deep learning taxonomy	Partial	Medium	Emphasis on models and pipelines rather than detailed vision analysis
Zhang et al.[12]	2024	Vision-language models	VLM-based taxonomy	Yes	Medium	Broad scope across vision tasks, not specific to video captioning
Qasim et al.[13]	2025	Dense video captioning	Task-based taxonomy	No	Low	Focus on temporal localization and dense captions, limited comparison of visual encoders
Liu and Song [14]	2025	Dense video captioning	Dataset and task-oriented taxonomy	No	Low	Emphasis on resources and datasets rather than model-level analysis
Li et al.[15]	2025	Large vision-language models	Transformer-based taxonomy	Yes	Medium	Focus on large models, limited discussion of classical approaches and comparisons
This Work	2025	Vision feature extraction in video captioning	CNN vs Transformer vs MLLMs + spatial/temporal/multimodal	Yes	High	-

2. Method

2.1. Study Selection Process

The review is semi-systematic to ensure that methodological rigor is maintained, as well as coverage of new developments in the area of video captioning. The search of the literature was performed with the help of several databases, such as Scopus, IEEE Xplore, Springer Nature Link, and arXiv. At the search phase, the filters were applied to select only open-access and English-language publications.

A total of 890 records were initially identified. Then we found out the duplicate papers. We found 180, and we removed them. And 20 papers we didn't consider for other reason. After removing we had 690 records for screening. During the title and abstract screening stage, 390 records were excluded based on relevance, we studied them but those were not full fill our goal. After that we only had 300 reports for full-text retrieval. Of these, 20 reports were not accessible, and 280 reports were assessed for eligibility. After applying the inclusion and exclusion criteria (like some researcher using different metrics

for evaluate their model or some are using unknown dataset), 97 studies were selected for final analysis. The study selection process is illustrated using a PRISMA flow diagram (Fig. 1).

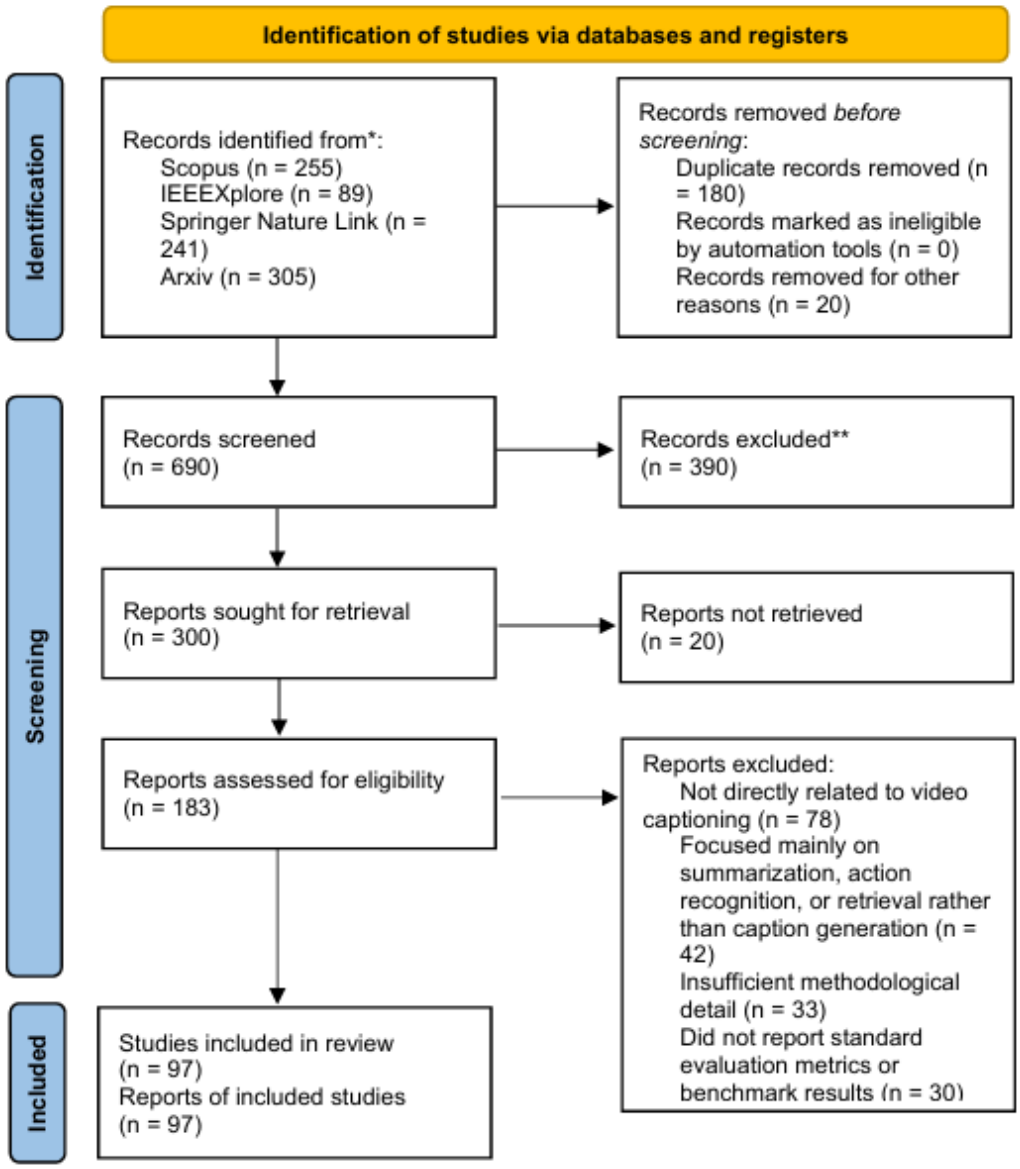


Fig. 1. PRISMA Flow Diagram of the Study Selection Process.

2.2. Inclusion and Exclusion Criteria

The studies were also incorporated in case they concentrated on automatic video captioning and offered adequate methodological and experimental information. The papers that covered related yet different tasks, like video summarization, action recognition, or video retrieval without caption generation, were avoided. Other exclusion criteria were the lack of technical detail, the absence of standard evaluation measures, and the research that was not in line with the scope of this review. A study was only included if:

- The study needed to be specifically on captioning videos (not summarization, not just action recognition [29]).
- It needed to describe in detail its vision feature extraction method.
- At least one standard set of captioning metrics (BLEU [30], METEOR [31], CIDEr [32], ROUGE [33]) had to be reported.

- It used a dataset that was publicly available.

We filtered out papers that were either only on the language-modeling side or were not in English or those that applied existing models without modification. After the process of screening full text, we had around 100 papers and these were the established baselines as well as more experimental state of the art systems.

2.3. Data Extraction and Organization

For each study, we documented some data:

- bibliographic information (authors, year, venue),
- the vision backbone (CNNs such as VGG, ResNet, C3D, I3D, or transformers like ViT, Swin, TimeSformer).
- datasets utilized (with special emphasis on ActivityNet, YouCook2, MSR-VTT, and MSVD, as they were the most frequent to be mentioned).
- evaluation metrics (BLEU, METEOR, ROUGE-L, CIDEr).
- the reported results for each dataset–metric pair.

This information was collected into a master table, which allowed us to compare methods without constantly going back to the original papers.

2.4. Limitations of the Review

Though numerous efforts were done to guarantee the thorough and systematic analysis, there are several restrictions that should be noted. The review contains only English-language and that paper which can be accessed in free (open-access publications) first. Which may cause selection bias and leave out other studies that were published in other languages or limited-access journals. Second, the addition of preprint papers of arXiv and peer-reviewed publications can bring variability to the quality and maturity of the methods analysed.

In addition, the chances of publication bias that can affect the results are also present because the results of studies that reported positive or improved results are more likely to be published. The heterogeneity of datasets and evaluation protocols applied in the studies is another important limitation. Variations in benchmark datasets, preprocessing methods, and evaluation measures complicate to compare methods directly and can impinge the applicability of the conclusions in general.

Finally, the review is a semi-systematic review; thus, it does not purport to cover all literature that is available especially due to the fast-paced development of the video captioning field.

2.5. Classification Framework

To facilitate the understanding of the impact of the visual encoder influences on the performance, the reviewed studies will be categorized according to the used visual feature extraction mechanisms instead of the type of language model used. This perception underscores the role played by visual encoding in development of downstream caption generation. Two main types are differentiated:

CNN-based methods: CNN-based models are based on 2D or 3D convolutional operations to identify spatial and short-term temporal features on video frames. Hybrid methods (using CNN as encoders and RNN as decoders) are variants.

Transformer-based methods: These models use self-attention mechanisms, such as vision transformers or spatio-temporal transformers. Can learn long-range dependencies and more intricate interactions in video sequences. Since there are numerous methods that can be classified under this classification, the methods that use transformer are further divided as follows:

- Vision Transformer-based models: These approaches use transformer encoders primarily for visual feature extraction, replacing traditional convolutional backbones.

- Spatio-temporal transformer models: These are the models which are an explicit model of the temporal relationship between the frames, using attention mechanisms in order to gain a better understanding of the sequence of motion and events.
- Multimodal Large Language Models (MLLM): These methods are based on extending transformer models to incorporate visual, audio and textual modalities within a single model. Which allows models to have complex reasoning.

This hierarchical structure is a reflection not only of the architectural development of the discipline but also of the growing focus on the time-based modeling and multimodal thinking. Although CNN-based ones were dominant before 2021, transformer-based models have taken over as they are much more capable of capturing long-range dependencies and cross-modal interactions.

2.6. Analytical Approach

We analyzed the selected studies in three ways:

- Popularity evolving over time: how the share of CNN-based vs. transformer-based approaches shifted, particularly before and after 2021.
- Usage of datasets: what were the most common datasets used, and what they told us about the kind of videos the researchers cared about (short videos, long unedited videos, instructional videos, etc.).
- Performance comparison: average metric scores across method categories and datasets, to highlight where CNNs or transformers tended to perform better or worse.

We also examined the metrics themselves. Others, such as BLEU, are convenient, but restricted to word overlap. These others, like CIDEr or METEOR, are more prone to capture semantic richness. Based on these observations, we make suggestions about which evaluation tools may be more reliable in future work.

3. Results and Discussion

3.1. Overview of Collected Literature

The final corpus analyzed in this review consists of 100 peer-reviewed studies published between 2017 and 2025. These works show how video captioning methods have evolved over the past decade, moving from convolutional vision encoders to transformer-based architectures. For every study, bibliographic details (Year), the specific vision feature extraction method (how each paper analyzed the vision part of the paper), the datasets employed (In Table 2 we represent the dataset as roman number, such as (Activity Net Caption [34]: i, YouCook2 [35]: ii, MSR-VTT [36]: iii, MSVD [37]: iv, ViTT: v, VATEX [38]: vi, Flickr30k [39]: vii, CharadesEgo [40]: viii, NoCaps [41]: ix, Ego4D [42]: x, Charades: xi, M-VAD [43]: xii).), and quantitative evaluation scores across standard metrics (Evaluation metrics we collect BLEU@4, METEOR, ROUGE-L, CIDEr) were recorded. This information was consolidated into a comprehensive master Table 2, which serves as both a reference for the breadth of literature covered and the empirical foundation for the subsequent trend, performance, and dataset analyses presented in this section.

Table 2. Summary of Reviewed Video Captioning Papers (2017–2025) with Vision Methods, Datasets, and Evaluation Results.

Year	Ref.	Vision Feature Extraction/Following Method	Method Type	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
2025	[44](i)		TB	4.57	2.55	8.65	13.62	12.82
	[45](i)		TB		0.13	0.14	0.15	
	[46](i)	CLIP	TB		2.68	9.05		34.92
	[46](ii)				2.04	6.53		36.09
	[47](i)	LLAVA-NEXT-7B	TB		6.9	13.2		15.6
	[47](iii)				17.1	23.0		24.0

Year	Ref.	Vision Feature Extraction/Following Method	Method Type	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
2024	[48](ii)	Visual encoder. The first step is to encode the video into features f = F(V), f ∈ RN ×D , where N is the feature resolution	TB		1.92	7.51		39.06
	[49](i)		TB			9.0		41.2
	[49](v)					4.0		18.5
	[49](ii)					3.6		15.4
	[49](i)					10.0		37.8
	[49](v)					5.8		25.2
	[49](ii)					7.1		32.9
	[50](iii)	MICap + CLIP-ViT + BERT	TB		48.0	29.4	63.7	52.1
	[50](vi)				30.5	23.6	43.5	40.9
	[51](iii)	CLIP	TB				27.9	49.87
	[51](iv)						31.46	63.51
	[28](iii)	ViT-G/14 (from EVA-CLIP)	TB		33.4			74.6
	[28](iv)				51.0			179.1
	[28](ii)				17.6			131.2
	[52](i)	CLIP	TB		2.38	8.55		33.01
	[52](ii)				1.63	6.08		31.66
	[53](x)		TB		0.284	0.387	0.589	1.738
	[54](vii)	LLAVA	TB		32.8	26.7	55.4	88.9
	[54](iii)				21.0	23.9	49.5	33.1
	[54](ix)				44.4	29.9	60.6	111.4
	[54](vi)				18.4	19.0	40.8	33.8
	[54](ii)				35.3	28.6	57.2	120.8
2023	[55](i)	CLIP	TB			17.0		28.0
	[55](ii)					24.0		50.1
	[56](i)				14.5	17.97	36.56	31.13
	[56](ii)				9.4	17.94	34.55	48.7
	[57](i)	C3D + multi scale - self attention+ TCL	TB		2.1	7.73		28.9
	[57](ii)				0.92	4.53		23.21
	[58](i)		TB		2.06	8.22		
	[59](i)		TB	3.75	1.87	8.33		
	[60](i)		TB	12.75		16.01		29.35
	[60](ii)				0.99	4.81		24.11
	[61](i)	C3D + TSP + I3D	CB	11.7		16.35		33.33
	[61](ii)				1.04	5.01		26.52
	[62](i)		TB	3.1	0.63	6.34		
	[63](i)	CNN + DETR	TB		1.88	8.61		29.97
	[64](i)	I3D	TB	4.13	1.8	10.54		30.62
	[65](i)	2D-CNN + 3D-CNN	CB	11.36		16.56		26.16
	[65](ii)		CB		7.78	16.56		33.92
	[66](i)	CLIP	TB			2.7		7.5
	[66](ii)					2.1		4.9
	[67](i)	I3D + C3D + VGGish	CB	4.03	1.91	10.24		32.82
	[68](i)	Audio - VGGish, Visual - I3D, Textual - Word Embeddings	CB	4.27	2.06	10.09		
	[69](i)	C3D	CB		1.08	9.05		27.92
	[70](i)		TB		0.98	4.85		23.87
	[70](ii)				2.14	8.17		31.26
	[71](iii)	I3D + C3D	CB		43.9	30.2	62.6	50.8
	[71](iv)				56.9	37.6	73.4	93.1

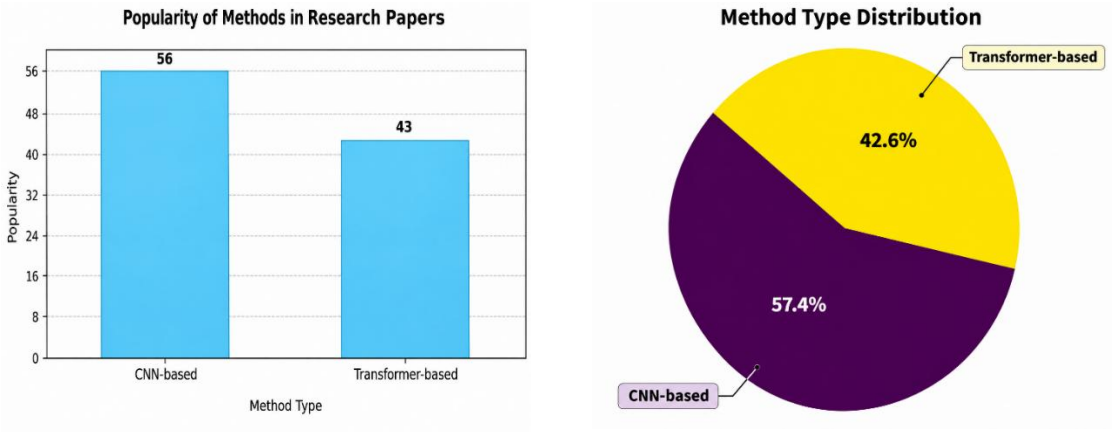
Year	Ref.	Vision Feature Extraction/Following Method	Method Type	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
2022	[72](ii)	CNN + Transformer	TB		0.87	11.99	23.25	0.86
	[73](i)	C3D and TSN	CB			11.01		
	[74](i)	CLIP	TB		13.38	17.48	35.0	31.29
	[75](i)	R(2+1)D and SA-TSP + I3D	CB			11.5		
	[76](i)	ResNet-200 + BNInception	TB		13.2	17.55	33.1	28.32
	[76](ii)			16.76	9.82	18.23	37.32	49.18
	[77](i)	VGGish + I3D + HOI + GloVe	CB	10.91	9.45	16.36		
	[78](iv)	EfficientNet B7 network	CB	77.8	75.0	34.7		
	[79](ii)		TB					
	[80](i)	ResNet200 + BN-Inception	CB	12.45		16.43		30.08
	[81](i)	Bi-SST + C + IIC (C3D features)	CB	4.01	1.91	7.89	19.2	11.01
	[82](i)		TB		10.31	15.78		23.72
	[83](i)	C3D	CB				9.25	
	[84](i)	C3D	CB	3.58	1.68	7.91		23.02
	[84](ii)				0.89	4.94		19.7
	[85](i)		TB		3.84	21.0	9.64	1.88
	[85](ii)						10.49	0.96
	[86](i)	2D-CNN	CB	5.58	3.89	12.89	25.37	26.52
	[86](ii)				1.09	4.31	10.51	9.07
	[87](iii)	InceptionResnetV2 + C3D + Faster-RCNN	CB		43.5	29.0	62.7	51.5
	[87](iv)				59.2	37.7	75.1	104.0
	[88](iii)	CLIP	TB		46.9	31.3	64.7	61.2
	[88](iv)				45.7	30.9	64.3	60.7
	[88](iii)				44.2	30.3	62.6	57.4
	[88](iv)				56.5	39.2	76.9	112.7
2021	[89](i)		TB	2.42	1.2	5.82		10.87
	[89](ii)					2.43		4.88
	[90](i)	C3D + TSN + Faster R-CNN	CB		2.84	9.74		28.09
	[91](i)	MART feature extraction (local & global)	TB		10.21	17.38		23.66
	[91](ii)				7.87	16.5		34.7
	[92](i)		TB	1.33	2.71	7.06		14.25
	[93](i)		TB			20.45	15.8	10.24
	[94](i)		TB	6.04	2.78	11.79		
	[95](i)	ResNet + C3D	CB			15.8		12.6
	[96](i)	ResNet + R3D	CB		9.45	16.36		19.4
	[97](i)		TB					
	[98](i)	C3D	CB	1.33	2.78	7.49	13.02	21.21
	[99](i)	ResNeXt-101 + ImageNet	CB			18.51	14.39	1.93
	[100](i)	R-CNN + Bi-RNN	CB	5.21	2.47	10.93	22.32	14.68
	[100](ii)				1.3	4.02	10.28	3.08
	[101](iv)	Inception-v4 + R(2+1)D + Faster-RCNN	CB		0.504	0.342	0.708	92.43
	[102](iii)		CB		40.8	28.3	60.8	49.5
	[102](iv)				52.8	35.5	72.9	94.3
	[103](xi)	residuals-assisted encoder (RAE) for video feature extraction	CB		16.5	18.0		21.0
	[103](iii)				42.8	29.3		48.7
	[104](iii)	CLIP	TB		47.2	31.2	64.8	60.0
	[105](iv)	VGG16, seq2seq(VGG16 (fc7 layer) for visual feature extraction,time-distributed FC layer, two bidirectional LSTM layers with Stacked Attention and Spatial Hard Pull)	CB	0.3961	0.1886			

Year	Ref.	Vision Feature Extraction/Following Method	Method Type	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
2020	[106](i)		CB		1.53	11.04	21.3	28.2
	[106](ii)				1.64	12.2	27.97	41.44
	[107](i)	C3D	CB			9.74		11.29
	[107](ii)					3.7		
	[108](i)	TSN	CB		2.85	11.49		49.34
	[109](i)	CNN Feature	CB	2.87	1.35	6.21		13.82
	[110](i)	Resnet200 + I3D	CB	2.91	5.32	11.28		14.03
	[111](i)	C3D+ I3D	CB	2.6	1.07	7.31		15.3
	[112](i)	I3D	CB	3.84	1.88	8.44		
	[113](v)	Compact 2D and S3D	CB		12.04	13.82	39.05	3.8
	[113](ii)		TB			12.43	33.1	0.9
	[114](iii)	R(2+1)D + T5 + finetune	CB		41.7	28.5		
	[114](ii)				5.3	13.4		
	[115](iii)	Encoder (CNN + LSTM)	CB	39.1	34.1	28.5		62.0
	[116](iii)	InceptionResNet-v2+ C3D	CB		39.9	27.5	59.9	40.3
	[116](iv)				47.2	34.3	70.9	72.1
2019	[117](i)	C3D	CB			10.71		31.41
	[118](i)	C3D + PCA	CB	2.63	5.19	9.54	22.98	15.81
	[119](i)	3D CNN + 3D Segment-of-Interest pooling	TB	1.63	4.06	8.81	19.63	19.88
	[120](i)	C3D	CB	0.82	1.69	4.78	9.6	10.53
	[121](i)	C3D	CB	2.94	0.93	8.82		30.68
	[122](i)	NaN	CB				9.9	
	[123](ii)	VFE Not addressed VideoBERT	TB	6.8	4.04	11.01	27.5	0.49
		VFE Not addressed VideoBERT + S3D	TB	7.59	4.33	11.94	28.8	0.55
	[124](iii)	Inception-Resnet-v2 CNN + LSTM	CB		0.404	0.27	0.61	0.483
	[124](iv)				0.48	0.336	0.705	0.865
	[125](iii)	ResNet-200	CB		41.4	28.2		46.9
	[125](iv)				56.9	36.2		90.6
	[126](iv)	EtENet + GoogLeNet	CB		48.9	32.4	78.0	69.4
	[126](iv)				50.0	34.3	86.6	70.2
2018	[127](i)	CNN	TB	4.47	2.23	9.56		
	[128](i)	CLIP	TB	5.84	13.52	14.75	25.88	
	[129](i)	3D CNN	CB	4.41	2.3	9.6	19.1	16.28
	[130](i)	C3D	CB	2.27	0.74	6.93		13.21
	[131](i)	GRU	CB	2.62	1.27	6.3	12.55	18.77
	[132](i)		CB		4.0	12.44		31.1
	[133](iii)	Inception-V4 + SA-LSTM + Reconstructors (Global)	CB		39.1	26.6		
	[133](iv)				52.3	34.1		
	[133](iii)	Inception-V4 + SA-LSTM + Reconstructors (Local)					59.3	42.5
	[133](iv)						69.8	80.3
	[134](iv)	Faster-RCNN + Bi-directional LSTM + LSTM with Attention	CB	0.554	0.438	0.326	0.693	0.522
	[135](xi)	CNN+Bi-LSTM	CB	29.4	18.8			
	[135](iii)				41.3	28.7	61.7	48.0
	[136](iv)	ResNet + C3D(CGN only (visual features only))	CB		26.5			26.4
	[136](iv)	ResNet + C3D(DSN + CGN (dynamic semantic features only))			35.8			

Year	Ref.	Vision Feature Extraction/Following Method	Method Type	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
	[136](iii)	ResNet + C3D (DSN + SSN + CGN (SeFLA))			41.8			60.1
	[136](iv)	ResNet + C3D (DSN + SSN + CGN (SeFLA))			50.0			94.3
	[136](iv)	ResNet + C3D (SSN + CGN)			41.4			
	[137](iv)	GoogLeNet and Inception-v4 (Multi-Task Video Captioning with Entailment Generation)	CB		54.4	35.9	72.7	88.0
		GoogLeNet and Inception-v4 (Multi-Task Video Captioning with Video Prediction)	CB		54.1	35.6	72.9	88.1
		GoogLeNet and Inception-v4 (Multi-Task Video Captioning with Video Prediction + Entailment Generation)	TB		52.6	35.0	71.5	84.4
2017	[34](i)	C3D	CB	8.45	5.52	10.03		29.92
	[138](iv)	C3D		56.7	46.9	31.4		
	[138](xii)	C3D + VGG	CB			6.9		
	[138](iv)	C3D + VGG		60.1	51.1	32.6		
	[139](iii)		CB		44.2	29.4	62.6	50.5
	[140](iv)	2-D/3-D CNN to extract visual features from selected video frames/clips, producing video representations through mean pooling over these visual features	CB	62.8	52.8	33.5		74.0

3.2. Popularity Trends in Vision Feature Extraction Methods

The year-by-year breakdown (Fig. 2(1a,1b) & Fig. 3): CNN-based vision models dominated the field in early period. But transformers have been steadily catching up and after 2021, overtaking them. Early papers almost universally relied on 2D CNN + temporal modeling (LSTM, GRU) [141] or 3D CNN variants, because at that time CNN was the best visual feature extraction model layer. Transformers were mostly used for encoding and decoding text. In total (Fig. 2(a)), across 2017 to 2025, 57.4% of the papers we reviewed used CNN-based mode for video analysis. In contrast, 42.6% of the papers used transformer-based methods. Before 2021, transformers barely registered in the vision feature extraction slot. From 2021 onward, the transformer line in the plot tilts sharply upward, crossing the CNN line and staying ahead.



(a) Popularity of Methods (b) Method Distribution

Fig. 2. Method popularity and distribution over (2017-2025).

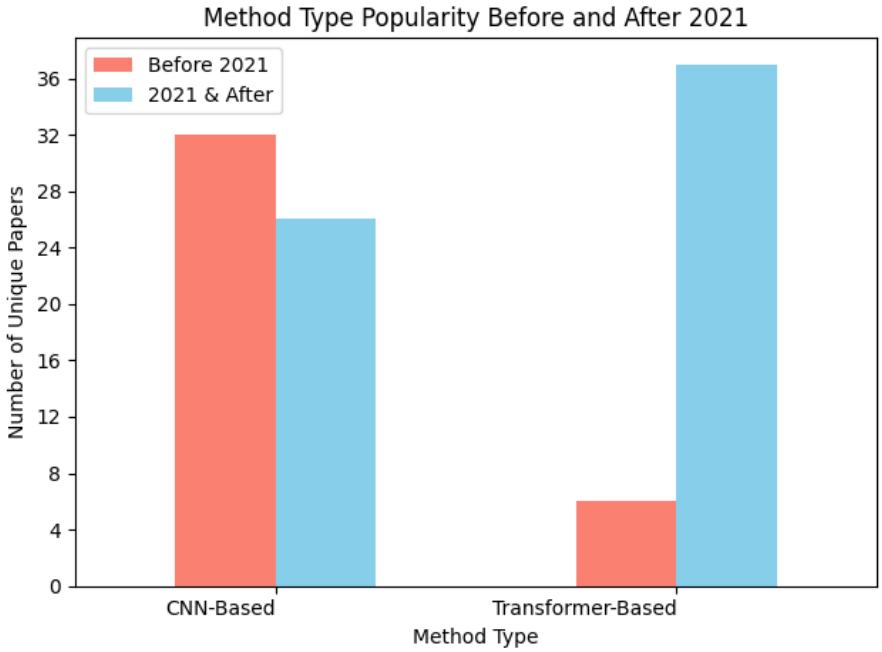


Fig. 3. Method trend before and after 2021.

This is not a coincidence. The multi-head self-attention mechanism has been around since 2017, but it wasn't until models like CLIP [142] appeared in 2021. CLIP shows that text and image alignment could work on a scale. That helps vision-language transformers gain serious traction. Nowadays hardware is advanced and there are many open sources, free to access large pretrained checkpoints are available online, which lowered the entry barrier. Video captioning researchers followed the broader computer vision shift toward transformer-based encoders.

Even though transformers became more popular after 2021, CNNs are still widely used in many research papers. This demonstrates that CNNs can still be utilized in some scenarios, especially when computing power is limited or when there isn't enough training data (video captioning data) for transformers. Transformers usually need strong hardware to compute and train with large datasets to perform well. On the other hand, CNNs can also be effective despite having fewer resources. So, while the shift toward transformers is clear, it's not a complete replacement. CNNs continue to play an important role, making the change more like a gradual handover than a full takeover.

3.3. Dataset Usage Patterns

The choice of dataset plays a critical role in the development and evaluation of video captioning models. Different datasets vary in size, domain coverage (like some datasets for action detection, some for only cooking item description), video length, annotation style, and degree of temporal segmentation. For video captioning, there are many datasets, but commonly used ones include ActivityNet, YouCook2, MSR-VTT, MSVD, ViTT, VATEX, Flickr30k, CharadesEgo, Ego4D, Charades, and M-VAD. In the papers we collected, most researchers like to use ActivityNet Captions, YouCook2, MSR-VTT, and MSVD datasets (Fig. 4).

Few researcher used multiple datasets to train their models. Some researchers combine the datasets so that the model can perform well. Table 3 summarizes four of the most widely used caption datasets in current research: ActivityNet Captions, YouCook2, MSR-VTT, and MSVD. ActivityNet Captions dataset is designed for dense video captioning. That means models must write several short descriptions for different parts of the same video. Built on the ActivityNet v1.3 dataset, it contains over 100,000 sentence descriptions paired with approximately 20,000 untrimmed YouTube videos. This dataset is particularly suited for tasks requiring an understanding of long-form video content and detecting transitions between semantically distinct events. It is widely used to test how well models can describe

videos and find the right time for each sentence. YouCook2 dataset focuses on the cooking domain. It comprises more than 2,000 long untrimmed instructional videos. Each video is segmented into a sequence of cooking steps with annotated temporal boundaries and step-specific captions. This makes it valuable for evaluating models' ability to handle procedural narratives and follow multi-step instructions. Due to its domain specificity, YouCook2 is often employed in research on recipe understanding and instructional video summarization. MSR-VTT dataset is a large-scale, general-purpose resource containing 10,000 short video clips sourced from the web. Each video is paired with 20 human-written captions. This variety helps the model learn to describe the same video in different ways. Covering a broad range of topics, from news and sports to music and movie trailers. This dataset serves as a robust baseline for evaluating caption generation in open-domain scenarios.

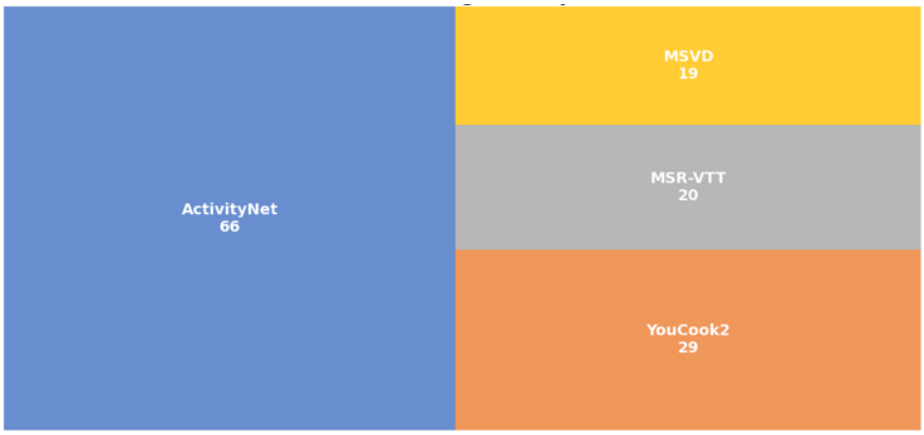


Fig. 4. Dataset Usage in Papers.

Table 3. Summary of Benchmark Datasets Commonly Used in Video Captioning.

Dataset	Video Length	Captions per Video	Temporal Segmentation	Domain	Primary Purpose
ActivityNet Captions [26]	~2 min	~3–4	Yes	General activities	Dense video captioning, temporal event localization
YouCook2 [27]	~5–10 min	~7–15	Yes	Cooking procedures	Procedural understanding, instructional video captioning
MSR-VTT [28]	10–30 sec	20	No	General web videos	General-purpose captioning, text-video retrieval
MSVD [29]	10–25 sec	~40	No	General web videos	Small-scale captioning, model prototyping, multilingual studies

MSVD (Microsoft Video Description Corpus) dataset is comparatively smaller, with 1,970 short YouTube videos. Each of them is accompanied by roughly 40 English captions and additional multilingual translations. Its smaller size makes it ideal for quick testing and experiments without heavy computing needs. Furthermore, its multilingual annotations have supported research in cross-lingual and multilingual video captioning. In Fig. 5, we plot the dataset video average length and the zone of the dataset. Here, the domain generality scale is 1 to 10. 1 means highly specific or limited zone; on the other hand, 10 means highly general. If the score is near 10, it means we can use this dataset for every sector.

The scatter plot highlights how different video captioning datasets vary in average video length and domain generality, guiding their optimal use. MSR-VTT and MSVD are short and cover many subjects, making them great for general tasks and quick testing. MSVD also has several languages. ActivityNet Captions has longer videos and more intensive annotations, which are better suited to tasks involving localization of sentences in time and comprehension of intricate event transitions. YouCook2, on the other hand, is most suitable for domain applications. The latter, e.g., instructional video summarization

and procedural reasoning, is due to its long-form cooking videos and lack of generality. The selection of the appropriate dataset is contingent on the scale of the task: MSR-VTT or MSVD are good general-purpose models, whereas the structured or instructional tasks require ActivityNet or YouCook2.

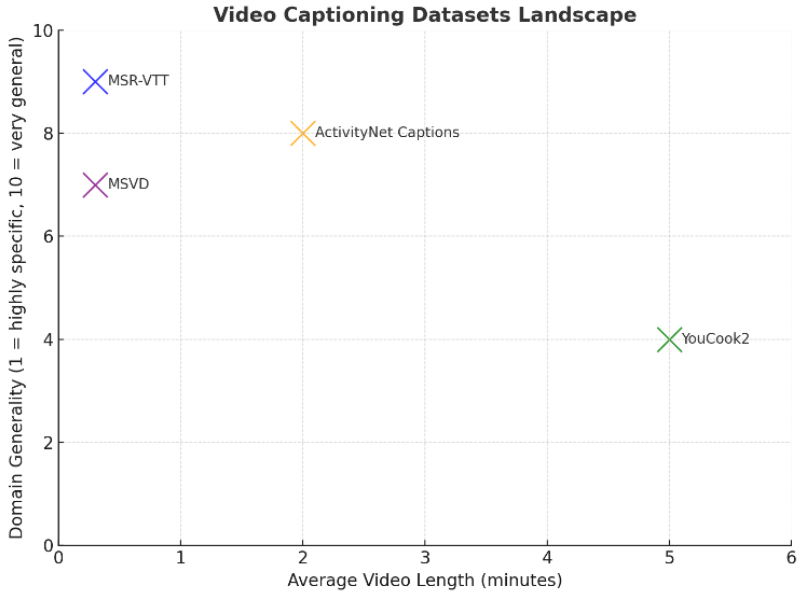


Fig. 5. Dataset Comparison by Average Video Length and Domain Generality.

3.4. Evaluation Metric Analysis

A video captioning system is not so easy to evaluate its effectiveness compared to a leaderboard (Table 4). Such common evaluation systems as BLEU, METEOR, ROUGE-L, and CIDEr were initially designed to handle machine translation or summarization. And they each have their own peculiarities when applied to image and video captions. BLEU is quick and replicable across studies, yet laser-focused on the exact n-gram matches that fail to consider the actual meaning. It implies it can reinforce robotic-sounding captions. METEOR attempts to fill in that gap by identifying synonyms and paraphrases, but requires lexical support, such as WordNet, and is slower. ROUGE-L is sensitive to sequence structure. It may be practical when the textual flow of the sentence is more important than the use of words, though it's still blind to meaning. CIDEr is designed for captioning and rewards rare, informative words, but is only effective when captions of good references have been made that can be compared.

Table 4. Summary of Benchmark Datasets Commonly Used in Video Captioning.

Metric	Primary Objective	Core Computation	Strengths	Limitations	Recommended Use in Video Captioning
BLEU (Bilingual Evaluation Understudy)	Checks how many words and short phrases in the candidate match the references, with a penalty for being too short.	Counts matching n-grams (1–4) and applies a brevity penalty.	Quick to compute, easy to compare across papers, well-established in benchmarks.	Overly focused on exact word matches; ignores synonyms and deeper meaning.	Large-scale benchmarking where speed and reproducibility matter more than nuance.
METEOR (Metric for Evaluation of Translation with Explicit Ordering)	Try to capture meaning, not just wording, by considering synonyms, stems, and paraphrases.	Uses a weighted precision–recall score with a penalty for word reordering.	Better correlation with human judgment than BLEU; recognizes	Slower to compute; depends on external lexical resources like WordNet.	Research tasks that value semantic accuracy and human-like language use.

Metric	Primary Objective	Core Computation	Strengths	Limitations	Recommended Use in Video Captioning
			alternate phrasings.		
ROUGE-L (Longest Common Subsequence)	Measures how similar the sequence of words is between candidate and reference.	Finds the longest common subsequence and normalizes by recall and precision.	Good at catching sentence-level structure and narrative flow.	Still surface-level; doesn't account for synonyms or meaning.	When preserving sentence structure is important, storytelling captions.
CIDEr (Consensus-based Image Description Evaluation)	Rewards captions that align with the consensus of multiple human-written references.	Uses TF–IDF–weighted n-gram similarity across all references.	Built specifically for captioning; favors informative, less common words; correlates well with human scoring.	Needs several high-quality references; scores drop with noisy or inconsistent data.	Main metric for assessing informativeness and overall alignment with human descriptions.

Comparing CNN-based and Transformer-based methods in these metrics (Fig. 6) we found a number of trends. In this case, it is possible to observe that transformer-based models were more likely to give a high median score on BLEU and CIDEr. What is accentuating their capability to portray bigger contexts and deliver more informative captions. Nonetheless, there was also heavier variance in their scores, indicating that they tend to perform well, but occasionally they may fail badly.

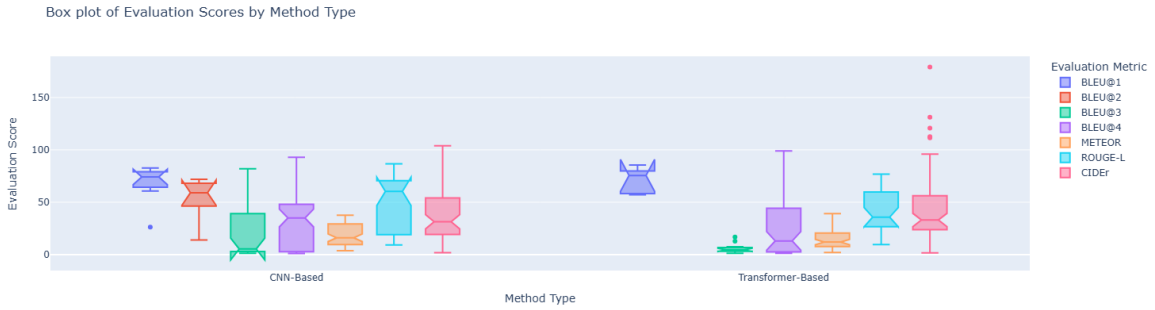


Fig. 6. Box plot comparing evaluation scores across CNN-based and Transformer-based methods using BLEU@1–4, METEOR, ROUGE-L, and CIDEr.

CNN-based models remained competitive in lower-order BLEU scores (BLEU-1, BLEU-2) but performed worse than those that consider more informative semantics, e.g., CIDEr and METEOR. That disparity is not that unexpected; Transformer is better able to capture long-range dependencies and incorporate context of other sections of a video. Simpler, surface-level descriptions still may be done using CNNs. Selecting the appropriate measures of evaluation and knowing how they interplay with the other model types is key to making an unfair but valuable evaluation in the video captioning study.

3.5. Comparative Performance by Method, Dataset

Fig. 7 shows a comparison of average evaluation scores between CNN-based and transformer-based methods. These four widely used datasets (ActivityNet Captions, MSR-VTT, MSVD, and YouCook2) are benchmark datasets used to show which types of models are doing well in all kinds of situations. Across all metrics, transformer-based models outperform CNN-based counterparts. This confirms transformer-based models' stronger ability to capture temporal dependencies and semantic relationships between image and text. The difference in performance can be best observed in the MSVD and YouCook2 datasets.

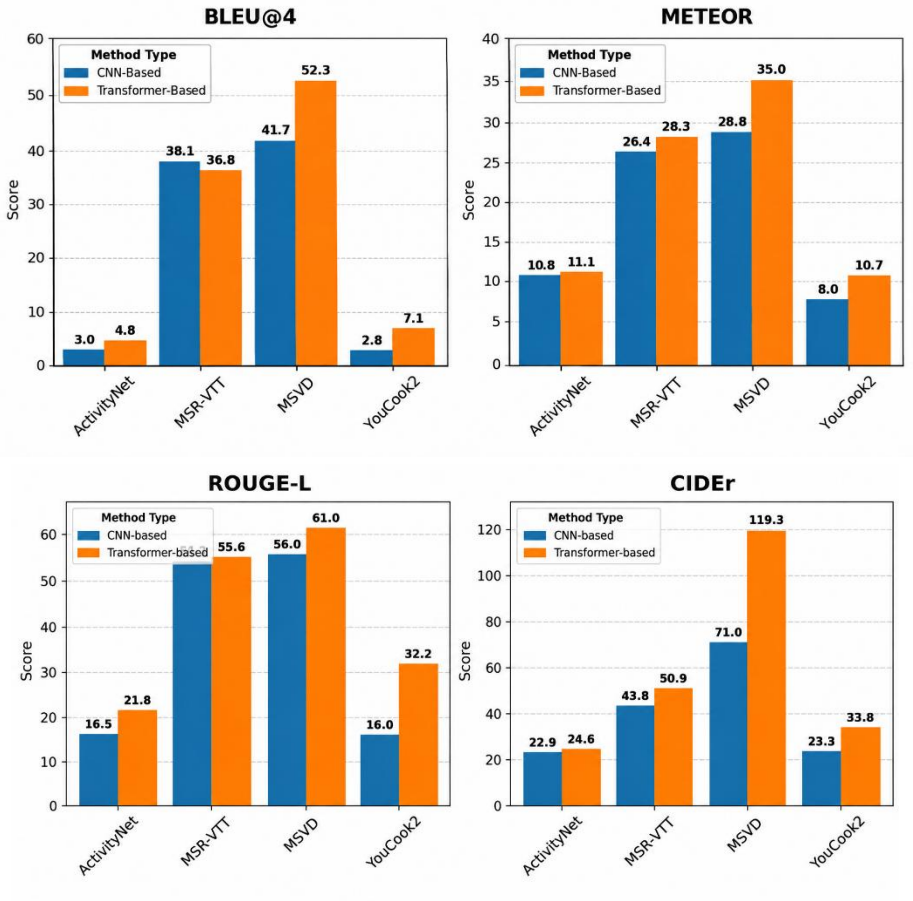


Fig. 7. Average evaluation scores for CNN-based and Transformer-based methods across four commonly used video captioning datasets.

MSVD consists of short video captioning. This dataset consists of well-structured video clips with several captions written by humans. Models are able to achieve high BLEU, METEOR, and CIDEr scores due to the large number of reference descriptions, which match a wider range of possible phrasings. YouCook2, on the other hand, contains lengthy videos with complex event sequences. The high performance of transformers on this dataset indicates that they are strong at modeling long-term temporal dynamics compared with CNN-based encoders, which are based on fixed-length feature aggregation.

However, this advantage should be interpreted with caution. Those datasets of shorter clips have more linguistic annotations, such as MSVD. On the other hand, lengthy video datasets such as ActivityNet Captions are more realistic in terms of their real-world characteristics, as events are longer-lasting and annotations are less common. Its captions are based on time, like subtitles, not like a full summary. Consequently, high scores on short and well-annotated datasets may not directly translate to similar accuracy in open-domain, real-world video captioning scenarios. The observations highlight the importance of balanced benchmarking that takes into consideration clip length as well as linguistic diversity in determining model performance.

4. Challenges and Insights

This part unites the key findings of the reviewed literature, outlines the main issues, existing gaps, the new tendencies, and potential future outlines of automatic video captioning.

4.1. Key Challenges

Although it has made tremendous improvements, there are still a number of challenges facing the development of video captioning systems. The representation of long-range temporal dependencies in video sequences is one of the key problems. Despite the enhanced temporal reasoning, transformer-based architectures are expensive in terms of computation and may not be as effective in ensuring the continuity of longer video stretches. Multimodal alignment is another important matter. Multimodal information integration, combining visual and textual messages and, in certain cases, audio, remains a challenge. A lot of models cannot represent small-grained associations between visual happenings and linguistic explanations. That is the reason why at times it generates the partial or not quite precise captions. Also, quality full dataset is another factor. Some datasets are not properly captioned; even if they caption or describe it properly, the number of captions for each video is not sufficient. Generalization across domains is another challenge. In case a dataset is built on the basis of a small number of actions or a single type of event, training a model with such a dataset will lead to a biased model. Moreover, the growing complexity of new architectures, especially transformer-based and MLLM models, that will be required to satisfy high computational demands and be challenging to implement such models in real-time or with limited resources.

4.2. Limitations of Existing Approaches

The critique also indicates that there are various constraints in current research practices. The characteristics of datasets greatly influence model performance. In case a model is trained on domain-specific data, it is hard to apply findings to other benchmarks. Popular evaluation measures, such as BLEU, METEOR, ROUGE, and CIDEr, are not without limitations. Such measures are not necessarily effective in capturing semantic meaning and contextual relevance. This causes differences between automatic assessment scores and human judgment. The other significant limitation is the absence of standardized evaluation protocols. The datasets, preprocessing procedures, and experimental design of various studies differ, and it is not easy to compare the findings of different studies fairly. Additionally, although MLLMs have shown promising capabilities, their integration into video captioning systems is still inconsistent, and their performance hasn't been thoroughly evaluated on standard benchmarks.

4.3. Trends and Insights

The study can be observed to have several significant trends. One of the significant changes is the transition from CNN-based feature extractors to transformer-based architectures, especially since 2021. Transformers have facilitated this change by their ability to better capture long-range dependencies and deal with complex multimodal interactions. Meanwhile, MLLMs also exhibit a strong movement towards cohesive vision-language systems that can generate captions more flexibly and context-awarely. At the same time, a temporal-aware model was also important. The other important point is that the performance of the models can frequently be highly dependent on the data used. These notes point to the increasing necessity of better guidelines in the choice of suitable models, depending on the requirements of the application, and the characteristics of the data.

4.4. Future Research Directions

Resting on the outlined challenges and trends, it is possible to determine numerous significant directions of further research. Creating low-cost, efficient techniques for captioning long videos. Another area of concern is the enhancement of multimodal fusion methods. Since it can be done more precise matching of visual, textual and audio information. Improving the strength and generalization ability of models, which are able to comprehend all domains (all types of events). And, how we will assess models, most important research. If the evaluation metrics are unable to evaluate all the aspects of the video, then it will be a disaster. Thus, more sophisticated assessment systems that are more indicative of semantic meaning and human judgment are urgently needed. Future research must also focus on the ways of enhancing the integration of multimodal large language models in video captioning systems. This involves putting them to the test on various datasets and in the field. Lastly, the development of light and resource-efficient model designs is also a significant issue, especially when it comes to real-time and edge deployment.

4.5. Practical Guidelines for Model Selection

According to our analysis, we can give several practical recommendations (Table 5) which could assist in the selection of the appropriate video captioning models in various situations. In the case of short videos, CNN-based models tend to perform well. They do well at deriving spatial features and cost significantly less in computing resources. Transformer-based models, on the other hand, can be used in longer or time-varying videos. They are preferable since they are able to represent long-range dependencies and temporal relationships in a better way.

Table 5. Guidelines for Model Selection in Video Captioning.

Application Scenario	Recommended Model Type
Short-duration video captioning tasks	CNN-based approaches
Long-video understanding and temporal reasoning	Transformer-based architectures
Multimodal reasoning and context-aware generation	Multimodal Large Language Models (MLLMs)
Real-time or resource-constrained deployment	Lightweight CNN-based models

Multimodal large language models (MLLMs) are more effective when the task demands more semantic processing or multimodal reasoning, e.g., context-sensitive captioning. Nevertheless, they are more computationally demanding and thus might not suit real-time applications or resource-constrained devices. Lightweight CNN-based models can also be used as an effective alternative in such situations. Thus, to choose a proper model, the complexity of tasks, the time requirements, and at-hand computational resources should be taken into account carefully.

5. Conclusion

In recent years, video captioning has evolved. It shifted to more dynamic deep learning systems, rather than the more rigorous CNN-based systems of template and retrieval. Older systems operated on CNN-RNNs. Next, there were Transformers and then multimodal large language models (MLLMs). An analysis of 100 publications of 2017-2025 indicates evident tendencies. Nevertheless, these tendencies are not necessarily as simple as they can be displayed in a chart. CNN-based architectures prevailed before 2021. They replaced Transformers, particularly when they were multimodal. This is not merely a change brought by a single architectural invention. Transformers are renowned for long-range interactions and subtle interactions between visual and textual inputs. The patterns of using databases have a different story. ActivityNet Captions, YouCook2, MSR-VTT, and MSVD continued to appear as the benchmarks of choice, but not all of them evaluate the same capabilities. Smaller datasets that are well captioned are likely to give inflated metric scores, whereas long-form video sets will always reveal the weakness in a model. High BLEU or CIDEr scores can also be readily interpreted as evidence of improvement, but measures based on n-gram overlap tend to favor superficial overlap over true semantic accuracy. In other cases, a model can be overheard to be so sure in describing something that was not occurring on screen. Despite this, recent years have brought momentum; there are still big challenges to overcome, like dealing with multiple minutes or hours of content without losing the thread, switching domains, and switching between languages (and doing so in a graceful manner), maintaining less biased datasets, and maintaining computational requirements at a lower level to be used more widely. Such aspects as hierarchical temporal modeling, retrieval-augmented generation, domain-aware pretraining, bias-aware dataset design, and more methods of evaluation that can measure what we actually care about in a caption are possible solutions. In perspective, it is reasonable to expect that the future generation of advancement will be built on a combination of trimmer Transformer-type frameworks, MLLMs and systems which are able to access external knowledge as it is required. Assuming these components can be coordinated, then maybe we will have captioning systems that not only pass the test but actually stand up in the slipstream of the real-world context, ranging from assisting somebody in a cooking tutorial in their own language to marking objectionable content on a social media platform. The speed or slowness with which

we arrive at it will be less due to the pursuit of the next metric score of higher rank and more to creating models that people themselves have confidence in to explain what they observe.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

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

[2] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, p. 11. Accessed: Feb. 17, 2026. [Online]. Available: https://arxiv.org/abs/1706.03762

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, Jun. 2021, p. 22. Accessed: Feb. 17, 2026. [Online]. Available: http://arxiv.org/abs/2010.11929

[4] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?," in *Proceedings of Machine Learning Research*, Jun. 2021, pp. 813–824. [Online]. Available: http://arxiv.org/abs/2102.05095

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *Proc. Mach. Learn. Res.*, vol. 202, pp. 20351–20383, Jun. 2023, [Online]. Available: http://arxiv.org/abs/2301.12597

[6] J.-B. Alayrac *et al.*, "Flamingo: A Visual Language Model for Few-Shot Learning," in *Advances in Neural Information Processing Systems 35*, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022, pp. 23716–23736. doi: 10.52202/068431-1723.

[7] B. Lin *et al.*, "Video-LLaVA: Learning United Visual Representation by Alignment Before Projection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2024, pp. 5971–5984. doi: 10.18653/v1/2024.emnlp-main.342.

[8] S. Islam, A. Dash, A. Seum, A. H. Raj, T. Hossain, and F. M. Shah, "Exploring Video Captioning Techniques: A Comprehensive Survey on Deep Learning Methods," *SN Comput. Sci.*, vol. 2, no. 2, p. 120, Apr. 2021, doi: 10.1007/s42979-021-00487-x.

[9] R. Akhare and S. Shinde, "Query Focused Video Summarization: A Review," in *Artificial Intelligence*, A. A. Sk, T. Turki, T. K. Ghosh, S. Joardar, and S. Barman, Eds., Cham: Springer Nature Switzerland, 2022, pp. 202–212. doi: 10.1007/978-3-031-22485-0_19.

[10] M. Liu, L. Nie, Y. Wang, M. Wang, and Y. Rui, "A Survey on Video Moment Localization," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–37, Sep. 2023, doi: 10.1145/3556537.

[11] M. Abdar *et al.*, "A Review of Deep Learning for Video Captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2024, doi: 10.1109/TPAMI.2024.3522295.

[12] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024, doi: 10.1109/TPAMI.2024.3369699.

[13] I. Qasim, A. Horsch, and D. Prasad, "Dense Video Captioning: A Survey of Techniques, Datasets and Evaluation Protocols," *ACM Comput. Surv.*, vol. 57, no. 6, pp. 1–36, Jun. 2025, doi: 10.1145/3712059.

[14] Z. Liu and R. Song, "Survey of Dense Video Captioning: Techniques, Resources, and Future Perspectives," *Appl. Sci.*, vol. 15, no. 9, p. 4990, Apr. 2025, doi: 10.3390/app15094990.

[15] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2025, pp. 1578–1597. doi: 10.1109/CVPRW67362.2025.00147.

[16] B. Günsel, A. M. Tekalp, and P. J. L. va. Beek, "Content-based access to video objects: Temporal Segmentation, visual summarization, and feature extraction," *Signal Processing*, vol. 66, no. 2, pp. 261–280, Apr. 1998, doi: 10.1016/S0165-1684(98)00010-3.

[17] L.-C. Chiu, T.-S. Chang, J.-Y. Chen, and N. Y.-C. Chang, "Fast SIFT Design for Real-Time Visual Feature Extraction," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3158–3167, Aug. 2013, doi: 10.1109/TIP.2013.2259841.

[18] Yuan Xie, Li-Feng Liu, Cui-Hua Li, and Yan-Yun Qu, "Unifying visual saliency with HOG feature learning for traffic sign detection," in *2009 IEEE Intelligent Vehicles Symposium*, IEEE, Jun. 2009, pp. 24–29. doi: 10.1109/IVS.2009.5164247.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, vol. 6, pp. 1–14, Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.

[23] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 4724–4733. doi: 10.1109/CVPR.2017.502.

[24] C. Yan *et al.*, "STAT: Spatial-Temporal Attention Mechanism for Video Captioning," *IEEE Trans. Multimed.*, vol. 22, no. 1, pp. 229–241, Jan. 2020, doi: 10.1109/TMM.2019.2924576.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[26] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[27] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Y. Feng and E. Lefever, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Dec. 2023, pp. 543–553. doi: 10.18653/v1/2023.emnlp-demo.49.

[28] B. He *et al.*, "MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2024, pp. 13504–13514. doi: 10.1109/CVPR52733.2024.01282.

[29] M. A. Rahim, A. S. M. Miah, H. S. Akash, J. Shin, M. I. Hossain, and M. N. Hossain, "An Advanced Deep Learning Based Three-Stream Hybrid Model for Dynamic Hand Gesture Recognition," vol. 1, pp. 1–26, Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.08035

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, in ACL '02. Morristown, NJ, USA: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[31] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the {ACL} Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds., Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909/

[32] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 4566–4575. doi: 10.1109/CVPR.2015.7299087w.

[33] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[34] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-Captioning Events in Videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 706–715. doi: 10.1109/ICCV.2017.83.

[35] L. Zhou, N. Louis, and J. J. Corso, "Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction," *Br. Mach. Vis. Conf. 2018, BMVC 2018*, vol. 1, pp. 1–16, Jul. 2018, [Online]. Available: http://arxiv.org/abs/1805.02834

[36] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 5288–5296. doi: 10.1109/CVPR.2016.571.

[37] D. Chen and W. Dolan, "Collecting Highly Parallel Data for Paraphrase Evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 190–200. [Online]. Available: https://aclanthology.org/P11-1020/

[38] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 4580–4590. doi: 10.1109/ICCV.2019.00468.

[39] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2641–2649. doi: 10.1109/ICCV.2015.303.

[40] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos," no. 2, pp. 1–3, Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.09626

[41] H. Agrawal *et al.*, "nocaps: novel object captioning at scale," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 8947–8956. doi: 10.1109/ICCV.2019.00904.

[42] K. Grauman *et al.*, "Ego4D: Around the World in 3,600 Hours of Egocentric Video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 11, pp. 9468–9509, Nov. 2025, doi: 10.1109/TPAMI.2024.3381075.

[43] S. Pini, M. Cornia, F. Bolelli, L. Baraldi, and R. Cucchiara, "M-VAD names: a dataset for video captioning with naming," *Multimed. Tools Appl.*, vol. 78, no. 10, pp. 14007–14027, May 2019, doi: 10.1007/s11042-018-7040-z.

[44] V. Estevam, R. Laroca, H. Pedrini, and D. Menotti, "Dense video captioning using unsupervised semantic information," *J. Vis. Commun. Image Represent.*, vol. 107, p. 104385, Mar. 2025, doi: 10.1016/j.jvcir.2024.104385.

[45] E. Blanco-Fernández, C. Gutiérrez-Álvarez, N. Nasri, S. Maldonado-Bascón, and R. J. López-Sastre, "Live video captioning," *Multimed. Tools Appl.*, vol. 84, no. 35, pp. 44863–44895, May 2025, doi: 10.1007/s11042-025-20908-w.

[46] Z. Xie, Y. Yang, Y. Yu, J. Wang, Y. Jiang, and X. Wu, "Exploring Temporal Event Cues for Dense Video Captioning in Cyclic Co-Learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 8, pp. 8771–8779, Apr. 2025, doi: 10.1609/aaai.v39i8.32948.

[47] S. Chu, S. Seo, and B. Han, "Fine-Grained Captioning of Long Videos through Scene Graph Consolidation," *Proc. Mach. Learn. Res.*, vol. 267, pp. 10801–10817, Jul. 2025, [Online]. Available: http://arxiv.org/abs/2502.16427

[48] T. Nishimura, A. Hashimoto, Y. Ushiku, H. Kameko, and S. Mori, "Recipe Generation from Unsegmented Cooking Videos," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 22, no. 9, pp. 1–25, Sep. 2026, doi: 10.1145/3649137.

[49] X. Zhou *et al.*, "Streaming Dense Video Captioning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2024, pp. 18243–18252. doi: 10.1109/CVPR52733.2024.01727.

[50] A. Hanna-Asaad, D. Aspandi-Latif, and T. Zaharia, "MI-Cap: A Multi-Modal Interpretable Model for Video Captioning," in *2025 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, Oct. 2025, pp. 1–8. doi: 10.1109/CBMI66578.2025.11339325.

[51] A. Nadeem, F. Sardari, R. Dawes, S. S. Husain, A. Hilton, and A. Mustafa, "NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative," *13th Int. Conf. Learn. Represent. ICLR 2025*, pp. 21619–21647, Feb. 2025, [Online]. Available: http://arxiv.org/abs/2406.06499

[52] M. Kim, H. B. Kim, J. Moon, J. Choi, and S. T. Kim, "Do You Remember? Dense Video Captioning with Cross-Modal Memory Retrieval," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2024, pp. 13894–13904. doi: 10.1109/CVPR52733.2024.01318.

[53] J. Xu *et al.*, "Retrieval-Augmented Egocentric Video Captioning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2024, pp. 13525–13536. doi: 10.1109/CVPR52733.2024.01284.

[54] W. Chai *et al.*, "AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark," *13th Int. Conf. Learn. Represent. ICLR 2025*, vol. 4, pp. 83337–83381, Apr. 2025, [Online]. Available: http://arxiv.org/abs/2410.03051

[55] A. Yang *et al.*, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 10714–10726. doi: 10.1109/CVPR52729.2023.01032.

[56] K. Yamazaki, K. Vo, Q. S. Truong, B. Raj, and N. Le, "VLTinT: Visual-Linguistic Transformer-in-Transformer for Coherent Video Paragraph Captioning," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 3, pp. 3081–3090, Jun. 2023, doi: 10.1609/aaai.v37i3.25412.

[57] T. Xu, Y. Cui, X. He, and C. Liu, "A latent topic-aware network for dense video captioning," *IET Comput. Vis.*, vol. 17, no. 7, pp. 795–803, Oct. 2023, doi: 10.1049/cvi2.12195.

[58] Y. Xie, J. Niu, Y. Zhang, and F. Ren, "Global-Shared Text Representation Based Multi-Stage Fusion Transformer Network for Multi-Modal Dense Video Captioning," *IEEE Trans. Multimed.*, vol. 26, pp. 3164–3179, 2024, doi: 10.1109/TMM.2023.3307972.

[59] Y. Wei, S. Yuan, M. Chen, and L. Wang, "Enhancing Multimodal Alignment with Momentum Augmentation for Dense Video Captioning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095659.

[60] Y. Wei *et al.*, "MPP-net: Multi-perspective perception network for dense video captioning," *Neurocomputing*, vol. 552, no. October, p. 126523, Oct. 2023, doi: 10.1016/j.neucom.2023.126523.

[61] T. Wang, J. Zhang, F. Zheng, W. Jiang, R. Cheng, and P. Luo, "Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos," vol. 2, pp. 1–1, May 2023, [Online]. Available: http://arxiv.org/abs/2303.06378

[62] J. Scherer, D. Bhowmik, and A. Scherp, "Event and Entity Extraction from Generated Video Captions," in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa, and E. Weippl, Eds., Cham: Springer Nature Switzerland, 2023, pp. 280–300. doi: 10.1007/978-3-031-40837-3_17.

[63] Y. Qian, Y. Mao, Z. Chen, C. Li, O. T. Bloh, and Q. Huang, "Dense video captioning based on local attention," *IET Image Process.*, vol. 17, no. 9, pp. 2673–2685, Jul. 2023, doi: 10.1049/ipr2.12819.

[64] L. H. Palivela, S. Swetha, M. Nithish Guhan, and M. Prasanna Venkatesh, "Dense Video Captioning Using Video-Audio Features and Topic Modeling Based on Caption," in *Proceedings of Fourth International Conference on Communication, Computing and Electronics Systems*, V. Bindhu, J. M. R. S. Tavares, and C. Vuppalapati, Eds., Singapore: Springer Nature Singapore, 2023, pp. 515–525. doi: 10.1007/978-981-19-7753-4_40.

[65] P. Li, P. Zhang, T. Wang, and H. Xiao, "Time–frequency recurrent transformer with diversity constraint for dense video captioning," *Inf. Process. Manag.*, vol. 60, no. 2, p. 103204, Mar. 2023, doi: 10.1016/j.ipm.2022.103204.

[66] Y. Jo, S. Lee, A. S. Lee, H. Lee, H. Oh, and M. Seo, "Zero-Shot Dense Video Captioning by Jointly Optimizing Text and Moment," vol. 1, pp. 1–19, Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.02682

[67] X. Huang, K.-H. Chan, W. Wu, H. Sheng, and W. Ke, "Fusion of Multi-Modal Features to Enhance Dense Video Caption," *Sensors*, vol. 23, no. 12, p. 5565, Jun. 2023, doi: 10.3390/s23125565.

[68] S. Han, J. Liu, J. Zhang, P. Gong, X. Zhang, and H. He, "Lightweight dense video captioning with cross-modal attention and knowledge-enhanced unbiased scene graph," *Complex Intell. Syst.*, vol. 9, no. 5, pp. 4995–5012, Oct. 2023, doi: 10.1007/s40747-023-00998-5.

[69] W. Choi, J. Chen, and J. Yoon, "Step by Step: A Gradual Approach for Dense Video Captioning," *IEEE Access*, vol. 11, pp. 51949–51959, 2023, doi: 10.1109/ACCESS.2023.3279816.

[70] W. Chen, J. Niu, and X. Liu, "MRCap: Multi-modal and Multi-level Relationship-based Dense Video Captioning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2023, pp. 2615–2620. doi: 10.1109/ICME55011.2023.00445.

[71] Z. Yan, Y. Chen, J. Song, and J. Zhu, "Multimodal feature fusion based on object relation for video captioning," *CAAI Trans. Intell. Technol.*, vol. 8, no. 1, pp. 247–259, Mar. 2023, doi: 10.1049/cit2.12071.

[72] W. Zhu, B. Pang, A. V. Thapliyal, W. Y. Wang, and R. Soricut, "End-to-end Dense Video Captioning as Sequence Generation," *Proc. - Int. Conf. Comput. Linguist. COLING*, vol. 29, no. 1, pp. 5651–5665, Sep. 2022, [Online]. Available: http://arxiv.org/abs/2204.08121

[73] Q. Zhang, Y. Song, and Q. Jin, "Unifying Event Detection and Captioning as Sequence Generation via Pre-training," in *Computer Vision -- ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 363–379. doi: 10.1007/978-3-031-20059-5_21.

[74] K. Yamazaki *et al.*, "VLCAP: Vision-Language with Contrastive Learning for Coherent Video Paragraph Captioning," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, Oct. 2022, pp. 3656–3661. doi: 10.1109/ICIP46576.2022.9897766.

[75] T. Wang, Z. Liu, F. Zheng, Z. Lu, R. Cheng, and P. Luo, "Semantic-Aware Pretraining for Dense Video Captioning," vol. 1, pp. 2019–2021, Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.07449

[76] J. Prudviraj, M. I. Reddy, C. Vishnu, and C. K. Mohan, "AAP-MIT: Attentive Atrous Pyramid Network and Memory Incorporated Transformer for Multisentence Video Description," *IEEE Trans. Image Process.*, vol. 31, pp. 5559–5569, 2022, doi: 10.1109/TIP.2022.3195643.

[77] A. Mittal, S. Ghosal, and R. Bansal, "SAVCHOI: Detecting Suspicious Activities using Dense Video Captioning with Human Object Interactions," vol. 2, pp. 1–14, Oct. 2022, [Online]. Available: http://arxiv.org/abs/2207.11838

[78] J. Madake, S. Bhatlawande, S. Purandare, S. Shilaskar, and Y. Nikhare, "Dense Video Captioning using BiLSTM Encoder," in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, May 2022, pp. 1–6. doi: 10.1109/INCET54531.2022.9824569.

[79] Y. Lu, Z. Zhang, Y. Chen, C. Yuan, B. Li, and W. Hu, "PIC 4th Challenge," in *Proceedings of the 4th on Person in Context Workshop*, in PIC '22. New York, NY, USA: ACM, Oct. 2022, pp. 17–21. doi: 10.1145/3552455.3555816.

[80] C. Liang, W. Wang, T. Zhou, and Y. Yang, "Visual Abductive Reasoning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 15544–15554. doi: 10.1109/CVPR52688.2022.01512.

[81] J. Jacob and V. . Devassia, "Dense Captioning of Videos using Feature Context Integrated Deep LSTM with Local Attention," in *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Nov. 2022, pp. 810–818. doi: 10.1109/I-SMAC55078.2022.9987416.

[82] Z. Guo, T.-J. J. Wang, and J. Laaksonen, "Post-Attention Modulator for Dense Video Captioning," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, Aug. 2022, pp. 1536–1542. doi: 10.1109/ICPR56361.2022.9956260.

[83] J. Dave and S. Padmavathi, "Hierarchical Language Modeling for Dense Video Captioning," in *Inventive Computation and Information Technologies*, S. Smys, V. E. Balas, and R. Palanisamy, Eds., Singapore: Springer Nature Singapore, 2022, pp. 421–431. doi: 10.1007/978-981-16-6723-7_32.

[84] W. Choi, J. Chen, and J. Yoon, "Parallel Pathway Dense Video Captioning With Deformable Transformer," *IEEE Access*, vol. 10, pp. 129899–129910, 2022, doi: 10.1109/ACCESS.2022.3228821.

[85] Z. Chang, D. Zhao, H. Chen, J. Li, and P. Liu, "Event-centric multi-modal fusion method for dense video captioning," *Neural Networks*, vol. 146, pp. 120–129, Feb. 2022, doi: 10.1016/j.neunet.2021.11.017.

[86] N. Aafaq, A. Mian, N. Akhtar, W. Liu, and M. Shah, "Dense Video Captioning With Early Linguistic Information Fusion," *IEEE Trans. Multimed.*, vol. 25, pp. 2309–2322, 2023, doi: 10.1109/TMM.2022.3146005.

[87] H. Ye, G. Li, Y. Qi, S. Wang, Q. Huang, and M.-H. Yang, "Hierarchical Modular Network for Video Captioning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 17918–17927. doi: 10.1109/CVPR52688.2022.01741.

[88] Z. Liu, X. Wu, and Y. Yu, "Multi-Task Video Captioning with a Stepwise Multimodal Encoder," *Electronics*, vol. 11, no. 17, p. 2639, Aug. 2022, doi: 10.3390/electronics11172639.

[89] [Z. Yu and N. Han, "Accelerated masked transformer for dense video captioning," *Neurocomputing*, vol. 445, pp. 72–80, Jul. 2021, doi: 10.1016/j.neucom.2021.03.026.

[90] M. Yu, H. Zheng, and Z. Liu, "Dense Video Captioning with Hierarchical Attention-Based Encoder-Decoder Networks," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9534164.

[91] X. Yan, Z. Fei, S. Wang, Q. Huang, and Q. Tian, "DVCFlow: Modeling Information Flow Towards Human-like Video Captioning," vol. 1, pp. 1–10, Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.10146

[92] B. Wu, G. Niu, J. Yu, X. Xiao, J. Zhang, and H. Wu, "Weakly Supervised Dense Video Captioning via Jointly Usage of Knowledge Distillation and Cross-modal Matching," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Z.-H. Zhou, Ed., California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 1157–1164. doi: 10.24963/ijcai.2021/160.

[93] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-End Dense Video Captioning with Parallel Decoding," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 6827–6837. doi: 10.1109/ICCV48922.2021.00677.

[94] C.-H. Lu and G.-Y. Fan, "Environment-Aware Dense Video Captioning for IoT-Enabled Edge Cameras," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4554–4564, Mar. 2022, doi: 10.1109/JIOT.2021.3104289.

[95] S. Lee and I. Kim, "DVC-Net: A deep neural network model for dense video captioning," *IET Comput. Vis.*, vol. 15, no. 1, pp. 12–23, Feb. 2021, doi: 10.1049/cvi2.12013.

[96] C. S. Kanani, S. Saha, and P. Bhattacharyya, "Global Object Proposals for Improving Multi-Sentence Video Descriptions," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2021, pp. 1–7. doi: 10.1109/IJCNN52387.2021.9533883.

[97] X. Dang, G. Wang, K. Xiong, and Z. Qin, "Dense Video Captioning for Incomplete Videos," in *Artificial Neural Networks and Machine Learning -- ICANN 2021*, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds., Cham: Springer International Publishing, 2021, pp. 665–676. doi: 10.1007/978-3-030-86383-8_53.

[98] S. Chen and Y.-G. Jiang, "Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2021, pp. 8421–8431. doi: 10.1109/CVPR46437.2021.00832.

[99] E. Boran, A. Erdem, N. Ikizler-Cinbis, E. Erdem, P. Madhyastha, and L. Specia, "Leveraging auxiliary image descriptions for dense video captioning," *Pattern Recognit. Lett.*, vol. 146, no. June, pp. 70–76, Jun. 2021, doi: 10.1016/j.patrec.2021.02.009.

[100] N. Aafaq, A. Mian, W. Liu, N. Akhtar, and M. Shah, "Cross-Domain Modality Fusion for Dense Video Captioning," *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 763–777, Oct. 2022, doi: 10.1109/TAI.2021.3134190.

[101] P. S *et al.*, "Exploration of Visual Features and their weighted-additive fusion for Video Captioning," vol. 1, pp. 1–6, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.05806

[102] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, "Semantic Grouping Network for Video Captioning," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2514–2522, May 2021, doi: 10.1609/aaai.v35i3.16353.

[103] M. Zhu, "Video Captioning in Compressed Video," in *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, IEEE, Jul. 2021, pp. 336–341. doi: 10.1109/ICIVC52351.2021.9526927.

[104] M. Tang, Z. Wang, Z. LIU, F. Rao, D. Li, and X. Li, "CLIP4Caption: CLIP for Video Caption," in *Proceedings of the 29th ACM International Conference on Multimedia*, in MM '21. New York, NY, USA: ACM, Oct. 2021, pp. 4858–4862. doi: 10.1145/3474085.3479207.

[105] M. M. Rahman, T. Abedin, K. S. S. Prottoy, A. Moshruba, and F. H. Siddiqui, "Video captioning with stacked attention and semantic hard pull," *PeerJ Comput. Sci.*, vol. 7, p. e664, Aug. 2021, doi: 10.7717/peerj-cs.664.

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

[107] T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu, "Event-Centric Hierarchical Representation for Dense Video Captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1890–1900, May 2021, doi: 10.1109/TCSVT.2020.3014606.

[108] T. Wang, H. Zheng, and M. Yu, "Dense-Captioning Events in Videos: SYSU Submission to ActivityNet Challenge 2020," vol. 2, pp. 3–6, Aug. 2020, [Online]. Available: http://arxiv.org/abs/2006.11693

[109] M. Suin and A. N. Rajagopalan, "An Efficient Framework for Dense Video Captioning," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 12039–12046, Apr. 2020, doi: 10.1609/aaai.v34i07.6881.

[110] Y. Song, S. Chen, Y. Zhao, and Q. Jin, "Team RUC_AIM3 Technical Report at Activitynet 2020 Task 2: Exploring Sequential Events Detection for Dense VidTeam RUC_AIM3 Technical Report at Activitynet 2020 Task 2: Exploring Sequential Events Detection for Dense Video Captioningeo Captioning," vol. 1, pp. 1–5, Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.07896

[111] V. Iashin and E. Rahtu, "Multi-modal Dense Video Captioning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 4117–4126. doi: 10.1109/CVPRW50498.2020.00487.

[112] V. Iashin and E. Rahtu, "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer," *Proc. Br. Mach. Vis. Conf. 2020*, pp. 1–16, Aug. 2020, doi: 10.5244/C.34.29.

[113] G. Huang, B. Pang, Z. Zhu, C. Rivera, and R. Soricut, "Multimodal Pretraining for Dense Video Captioning," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight, and H. Wu, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Dec. 2020, pp. 470–490. doi: 10.18653/v1/2020.aacl-main.48.

[114] B. Korbar, F. Petroni, R. Girdhar, and L. Torresani, "Video Understanding as Machine Translation," vol. 2, Sep. 2020, [Online]. Available: https://arxiv.org/abs/2006.07203

[115] Z. Fang, T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "Video2Commonsense: Generating Commonsense Descriptions to Enrich Video Captioning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 840–860. doi: 10.18653/v1/2020.emnlp-main.61.

[116] J. Hou, Y. Jia, X. Wu, and Y. Qi, "Video Captioning Using Weak Annotation," vol. 1, pp. 1–10, Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.01067

[117] Z. Zhang, D. Xu, W. Ouyang, and C. Tan, "Show, Tell and Summarize: Dense Video Captioning Using Visual Cue Aided Sentence Summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3130–3139, Sep. 2020, doi: 10.1109/TCSVT.2019.2936526.

[118] S. Yang, L. Li, S. Wang, D. Meng, Q. Huang, and Q. Tian, "Structured Stochastic Recurrent Network for Linguistic Video Prediction," in *Proceedings of the 27th ACM International Conference on Multimedia*, in MM '19. New York, NY, USA: ACM, Oct. 2019, pp. 21–29. doi: 10.1145/3343031.3350859.

[119] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko, "Joint Event Detection and Description in Continuous Video Streams," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Jan. 2019, pp. 396–405. doi: 10.1109/WACV.2019.00048.

[120] T. Rahman, B. Xu, and L. Sigal, "Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 8907–8916. doi: 10.1109/ICCV.2019.00900.

[121] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined Dense Video Captioning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 6581–6590. doi: 10.1109/CVPR.2019.00675.

[122] S. Chen *et al.*, "Activitynet 2019 Task 3: Exploring Contexts for Dense Captioning Events in Videos," vol. 1, pp. 1–6, Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.05092

[123] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A Joint Model for Video and Language Representation Learning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 7463–7472. doi: 10.1109/ICCV.2019.00756.

[124] L. Li and B. Gong, "End-to-End Video Captioning With Multitask Reinforcement Learning," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Jan. 2019, pp. 339–348. doi: 10.1109/WACV.2019.00042.

[125] J. Zhang and Y. Peng, "Object-Aware Aggregation With Bidirectional Temporal Graph for Video Captioning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 8319–8328. doi: 10.1109/CVPR.2019.00852.

[126] S. Olivastri, G. Singh, and F. Cuzzolin, "End-to-End Video Captioning," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Oct. 2019, pp. 1474–1482. doi: 10.1109/ICCVW.2019.00185.

[127] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-End Dense Video Captioning with Masked Transformer," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 8739–8748. doi: 10.1109/CVPR.2018.00911.

[128] Y. Xiong, B. Dai, and D. Lin, "Move Forward and Tell: A Progressive Generator of Video Descriptions," in *Computer Vision -- ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 489–505. doi: 10.1007/978-3-030-01252-6_29.

[129] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 7190–7198. doi: 10.1109/CVPR.2018.00751.

[130] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly Localizing and Describing Events for Dense Video Captioning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 7492–7500. doi: 10.1109/CVPR.2018.00782.

[131] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang, "Weakly Supervised Dense Event Captioning in Videos," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Nips, pp. 3059–3069, Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.03849

[132] S. Chen, Y. Song, Y. Zhao, J. Qiu, Q. Jin, and A. Hauptmann, "RUC+CMU: System Report for Dense Captioning Events in Videos," vol. 1, pp. 1–4, Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.08854

[133] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction Network for Video Captioning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 7622–7631. doi: 10.1109/CVPR.2018.00795.

[134] B. Zhao, X. Li, and X. Lu, "Video Captioning with Tube Features," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 1177–1183. doi: 10.24963/ijcai.2018/164.

[135] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video Captioning via Hierarchical Reinforcement Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 4213–4222. doi: 10.1109/CVPR.2018.00443.

[136] Sujin Lee and Incheol Kim, "Video Captioning with Visual and Semantic Features," *J. Inf. Process. Syst.*, vol. 14, no. 6, pp. 1318–1330, 2018, [Online]. Available: https://jips-k.org/digital-library/2018/14/6/1318

[137] R. Pasunuru and M. Bansal, "Reinforced Video Captioning with Entailment Rewards," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Sep. 2017, pp. 979–985. doi: 10.18653/v1/D17-1103.

[138] Y. Liu, X. Li, and Z. Shi, "Video Captioning with Listwise Supervision," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, pp. 4197–4203, Feb. 2017, doi: 10.1609/aaai.v31i1.11239.

[139] Z. Shen *et al.*, "Weakly Supervised Dense Video Captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 5159–5167. doi: 10.1109/CVPR.2017.548.

[140] Y. Pan, T. Yao, H. Li, and T. Mei, "Video Captioning with Transferred Semantic Attributes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 984–992. doi: 10.1109/CVPR.2017.111.

[141] H. S. Akash, M. A. Rahim, A. S. M. Miah, H.-S. Lee, S.-W. Jang, and J. Shin, "Two-Stream Modality-Based Deep Learning Approach for Enhanced Two-Person Human Interaction Recognition in Videos," *Sensors*, vol. 24, no. 21, p. 7077, Nov. 2024, doi: 10.3390/s24217077.

[142] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. Mach. Learn. Res.*, vol. 139, pp. 8748–8763, Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00020