IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 18, 2021, accepted August 25, 2021, date of publication August 27, 2021, date of current version September 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108565

Video Description: Datasets & Evaluation Metrics

MUHAMMAD RAFIQ, GHAZALA RAFIQ, AND GYU SANG CHOI, (Member, IEEE)
Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si 38541, South Korea

Corresponding author: Gyu Sang Choi (castchoi@ynu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2019R1A2C1006159 and Grant NRF-2021R1A6A1A03039493.

ABSTRACT Rapid expansion and the novel phenomenon of deep learning have manifested a variety of proposals and concerns in the area of video description, particularly in the recent past. Automatic event localization and textual alternatives generation for the complex and diverse visual data supplied in a video can be articulated as video description, bridging the two leading realms of computer vision and natural language processing. Several sequence-to-sequence algorithms are being proposed by splitting the task into two segments, namely encoding, i.e., getting and learning the insights of the visual representations, and decoding, i.e., transforming the learned representations to a sequence of words, one at a time. Implemented deep learning approaches have gained a lot of recognition for the reason of their superior computing capabilities and tremendous performance. However, the accomplishment of these algorithms strongly depends on the nature, diversity, and amount of data they are trained, validated and tested on. Techniques applied on insufficient and inadequate train/test data cannot deliver promising conclusions, consequently making it complicated to evaluate the quality of generated results. This survey focuses explicitly on the benchmark datasets, and evaluation metrics developed and deployed for video description tasks and their capabilities and limitations. Finally, we concluded with the need for essential enhancements and encouraging research directions on the topic.

INDEX TERMS Datasets, evaluation metrics, sequence to sequence, video description, video captioning, vision to language, vision to text.

I. INTRODUCTION

The rapidly growing digital culture has fascinated people to interact with thrilling multimedia data, i.e., images, videos, voice notes, and texts everywhere. Video, the globally renowned document type, is common everywhere at memorable events, instructional purposes, evidence apprehension, information exchange, business marketing. On average, there are more cameras, i.e., CCTV, digital recorders, or phone cameras, than the people on the face of the earth. According to CISCO annual internet report, [1], ''video will be 78% of mobile data traffic by 2021''. Visual data exploded to a drastic degree within a couple of recent years. As most of the data on the internet is visual data, robust algorithms are required to deal with this data efficiently. Nevertheless, for the machines to understand that visual data and automatically generate its precise interpretation is a big challenge. More than 500 hours of videos are uploaded to YouTube every minute [2] and while these numbers are staggering, watching all these videos are almost impossible. 24/7 operating video-streaming and sharing platforms are dealing effectively with the massive number of videos indexing and retrieval. Although these sites categorize videos based on their genre and duration, even though, instead of the entire long-duration video, a precise textual alternate will be more effective to serve the purpose and save time.

Configuration of natural intelligence, a compendium of general knowledge and prior life observations and experiments have constituted human beings highbrowed enough to presume and expound the plot of a scene or a tiny fragment of it in a single glance but machines, no doubt, need to acquire that astuteness for scene pertinent conception and appropriate expression. Adequate understanding and learning of visual data and then accurately describing it using a single natural language sentence can be categorized as captioning whereas multi-sentence or paragraph like captioning can be termed description. The development of video description has attracted substantial attention from researchers in

The associate editor coordinating the review of this manuscript and approving it for publication was Byung Cheol Song.

the inter-disciplinary field of computer vision exploring the diverse areas of physics, biology, psychology for optics understanding, image formation, visual information processing, and natural language processing for automatic caption generation.

The irrefutable benefits and real-time applications of video description have drawn considerable attention from experts and motivated them further to develop more technologically sound systems for the purpose. These applications include human-robot collaboration, efficient content search and retrieval, video surveillance, combined with speech can describe the graphical content to the visually impaired, automatic video subtitling, conversion of sign language videos into natural language, procedure generation for instructional videos, and autonomous vehicles. After meticulous research on video tagging and image captioning, the field of video description is in focus since the recent past. Continuous, thorough exploration of the field is improving the quality of generated captions day by day.

The research in captioning is gradually making progress in getting closer to human annotations and generating captions as analogous as possible to the human description. However, there are a few challenges faced while automatically describing a video using natural language sentences. One major challenge faced is the semantic gap or the visual ambiguity present in the supplied visual data. Identification of visual contents along with their spatial, temporal relationship, and interconnections is sometimes challenging because of the hard in capturing and expressing visual details. For humans, straightforward to interpret visuals, scenes, or gestures becomes tricky and challenging for the algorithm to comprehend. Objects and action detection mechanisms establish the basis of captioning systems. However, these mechanisms not contemplating audio and motion multimodal responses lack in expressing the diverse and complex scenes. Likewise, some visuals can only be interpreted accurately when there is an injection of general knowledge into the system for scene awareness. The perfect assessment of concise, efficient, and diverse caption generation plays an essential role in further enhancement and improvement. Most of the evaluation metrics used for the assessment and valuation are not task-specific and cannot guarantee success. Similar is the case with benchmark datasets. No task-specific standardized datasets exist for the assessment of algorithms, particularly concerning dense video captioning.

A. CLASSICAL APPROACH

Video description journey of evolution pioneered by the classical methods, based on SVO (Subject, verb, object) used to describe a visual [3]–[6], [7]. It is a two-stage pipeline where after identification of the subject, verb, and object in a series of frames, plugging in a pre-defined standard template is performed. Following vast exploration, the major limitation of classical methods is their dependence on fixed predefined templates, which cannot generate semantically rich natural language sentences, hence not analogous to human

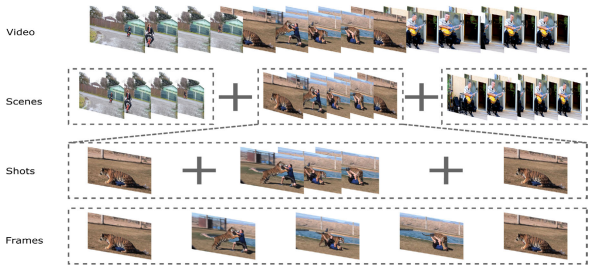


FIGURE 1. Hierarchical Structure of a Video.

annotations. Another limitation is the requirement of a set of objects and actions for recognition because of individual classifier training for identification. Furthermore, these methods can be considered useful in the scenarios where the video clip is short, or the number of objects or actions in that clip is limited, but not otherwise.

The temporal structure of the video is intrinsically layered [8]. The hierarchical structure of video comprises of the scene, shot, and frames, increasing in granularity, decreasing in semantics from top to bottom [9]. A single frame is the least logical unit of the video and represents a static image. The collection of such frames representing a solo camera motion shape a shot, and multiple consistent and coherent shots contributing to the matching concept or site produce a scene. This assortment of scenes constitutes a video, as shown in Figure 1. Segmentation of videos into scenes, shots, and frames facilitates efficient searching and indexing. Keyframe selection also contributes towards the reduction of computational cost and effective processing for summarization or description purposes. Different mechanisms to select a keyframe are available in the literature, like every fifth frame or every 16th frame or shot boundary detection mechanism. It significantly reduces the redundant frames to lower the computational cost but ensuring that none of the critical information is lost.

B. IMAGE CAPTIONING VS VIDEO CAPTIONING

Initial research on captioning used metadata for tagging [10]. Video tags are generally the names of items, actions, or activities in the video, which are the video's significantly considerable entities or events. The systems based on templates and probabilistic graphical models following the SVO approach were developed in the past to generate captions [11]. Image captioning extended the tagging process by adding a spatial relationship between the objects with the help of natural language modeling. The crucial challenge faced during caption generation is the semantic gap [11]. The exciting part of video captioning is to add information on the temporal relationship of the objects, actions, or events detected in the video and their temporal order while generating the captions. So, the difference to mention is, for image captioning, only spatial relationship is required, whereas, for videos to be described appropriately, spatial-temporal relationship [12], as well as temporal order, is mandatory for adequate description. High temporal dependencies, complex

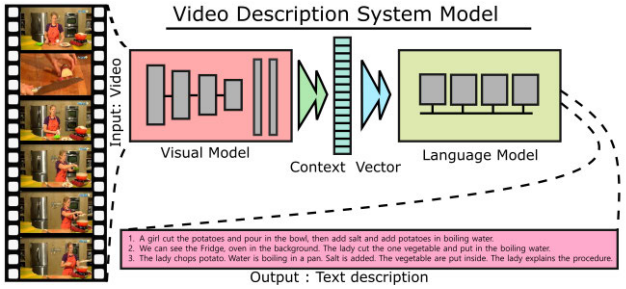


FIGURE 2. System Model for Video Description (Video frames and reference captions taken from MSVD dataset).

nature of the video, and diverse objects, scenes, actions, and inter-connectivity make it very difficult to accurately caption a video. However, recent deep-learning-based approaches and techniques for image captioning [13]–[21] and video captioning [22]–[29], [30]–[33] are noticeably pushing forward the research in this field.

Encouraged by the accomplishment of image captioning and machine translation tasks, the video captioning approaches primarily employ encoder-decoder architecture. The encoder-decoder framework is a neural network design configuration. The architecture is partition into two components, namely the encoder and the decoder. It has proven to be a cutting-edge machine translation technology. The research community around the globe has employed the modern approach to solving sophisticated tasks, i.e., image captioning, video description, text and video summarization [34], and visual question answering system/ conversational modeling, learning to execute, and movement classification. The encoder or visual model encodes a video by extracting visual information and generating a fixed dimension feature vector or context/ thought vector for use by the decoder or language model. The decoder processes the encoder-generated context vector for caption generation and generates one word at a time. Convolution neural networks (CNNs), recurrent neural networks (RNNs) and their variants, long short term memory (LSTM), gated recurrent unit (GRU) are used as visual and language models. In recent literature, various strategies are investigated for the quality enhancement and optimization of generated captions, including attention mechanism incorporation, hierarchical approaches, multimodal techniques, reinforcement learning integration, and transformer mechanism. Employment of these techniques boosted the performance of captioning systems, but the desire to get a human-like precise and accurate captioning for a supplied video is still intensively under consideration.

C. DENSE VIDEO CAPTIONING

Automatic event localization and textual alternatives generation for the complex and diverse visual data supplied in a video can be articulated as dense video captioning or video description. Dense video captioning is similar to dense image captioning [35] which localizes regions in image space and then describes those localized regions using natural language. For long videos description [36], proposal-module identifies long and short events based on the extracted features. Then each proposal consisting of unique start and end time accompanied by a hidden representation is fed to a language model for caption generation for each event leveraging context from neighboring events. The multi-scale event detection approach is used to overcome the limitation of overlapping events occurring in a video and introduced context utilization for related events caption generation.

D. VIDEO TO VIDEO SUMMARIZATION

In the current world, the role of multimedia for information exchange is beyond doubt. Videos are considered the best way to convey information, but storage requirements and time consumption for specific content retrieval make it inconvenient in certain circumstances. As a solution, the automatic video to video summarization technique produces a condensed version of a full-length video stream by extracting the most important content. Since the automatic mining of video semantic contents is complex, video metadata is mainly considered for content description and summary generation. [37] explored duplicate frame removal and stroboscopic imaging for generalized video summarization. Specific to user diverse preferences and expectations, [38], [39] proposed personalized summarization exploring incapability of inefficient generalized technique for resolving the specific individual requirements. Likewise, [40] also evaluated the degree of importance based on user behavior to carry out summarization reflecting the diversity of user choices and interests. Real-time [40] or live videos quick and instantaneous summarization is also explored by analyzing intrinsic video data and corresponding extrinsic metadata of the video stream. Sports videos highlight generation from a sports TV broadcast is also studied.

In this paper, a detailed exploration of benchmark datasets and evaluation metrics for assessing open-domain video description tasks is carried out. Benchmark datasets with their key attributes and train/ validation/ test split are presented, supporting their technical worth from the literature by qualitative & quantitative comparison among different models proposed with time. Evaluation metrics used to examine the quality of generated captions and their domain, computation concept, and limitation are investigated in detail. Finally, we identified future research directions for further improvement in the video description system.

1) PROBLEM STATEMENT-VIDEO DESCRIPTION

Suppose for a given video V, such that $V = \{v_1, v_2, \ldots .v_N\}$, with N frames or clips, a textual description comprising of automatically generated natural language sentence S where $S = \{w_1, w_2, \ldots .w_n\}$ consisting of n words, is required. Further, for dense or paragraph description, we need to generate a paragraph P, collection of temporally localized sentences, such that $P = \{S_1, S_2, \ldots ..S_T\}$ comprising of T sentences and each sentence stamped with its start and end time.

This paper is organized as follows: Section II provides a brief overview of the available literature on the topic, Section-III explores the benchmark datasets available for video description followed by Section-IV presenting critical analysis regarding these benchmark datasets. Section-V discusses in detail the metrics used for the evaluation of the video description system accompanied by Section-VI investigating the limitations, reliability, and improvements in the available metrics for appropriate evaluation. Section-VII elaborates the establishment of cross-modal models,i.e., visual and language, employing pre-training techniques for performance enhancement and Section-VIII compares the quantitative as well as the qualitative benchmark results, and at the end, the survey is concluded in Section-IX with few future directions.

II. LITERATURE REVIEW

Intensifying technological demands for automatic visual details understanding and content summarization has motivated the researchers to accomplish such capabilities better. The promotion and efficiency of neural networks have fashioned foremost advancements in accurately describing videos/ images and has drawn increasing attention, so become one of the hot research topics in the AI community around the globe. The available literature explores many aspects of the systems designed for the purpose. A majority of the available surveys [41]–[43] report the diverse aspects, along with their achievements and limitations. In the survey [28], the authors argued the visual to text transformation techniques targeting traditional/ classical natural language generation models as well as deep learning-based techniques for both image and video description. Considering classification and detection as necessary steps for action recognition, complex human action recognition and appropriate textual replacement are still challenging. The survey focuses on the challenges of fine-grained natural description, Intermediate representation learning, recounting of visual content, along with benchmark datasets and evaluation metrics. Authors presented natural language generation methods with recent advancements in image and video captioning. In survey [8], [44], and [45] authors explored the methodologies, datasets and evaluation metrics up to a certain extent.

Survey [11] aims at addressing the issues associated with caption generation, i.e., semantic supervision, mitigation of objective mismatch, dense captioning for jointly describing multiple events in a video, and their localization similarly in [41], the authors emphasized the algorithmic essence of different attention mechanisms and their application on the image captioning deep learning models available in the literature along with standard datasets and metrics requirement for the evaluation of the model.

III. DATASETS

Describing video is a much more challenging and computationally expensive task as compared to that image captioning. Proper understanding and precise interpretation of temporal relationships along with temporal order are accommodated while describing videos. A repository of video clips with its corresponding single or multiple annotation or description is referred to as a dataset that can work as a base for training, validating, and testing proposed models. Domain-specific datasets belong to the cooking, movies, social-media, wild, and human-action domains. Whereas open-domain datasets deal with a wide variety of videos, i.e., music, people, gaming, sports, news, education, vehicles, beauty, advertisement. Table-1 lists down a brief overview of the key attributes of these datasets and table-2 shows the training/validation/test splits of these datasets. Widespread benchmark datasets employed in recent research for video descriptions are elaborated as follows.

1. A man getting face painted by a woman artist.
2. A girl applied eye makeup to a guy.
3. A lady decorates a man's face.
4. A woman applies Joker makeup to a man's face.
5. A woman is applying makeup around a man's eye.

FIGURE 3. Example video frames and captions from MSVD dataset.

A. MSVD (MICROSOFT VIDEO DESCRIPTION)

MSVD dataset [46] is one of the earlier available and frequently used corpora by the research community around the globe. It is a collection of 1970 YouTube video clips provided with human annotations. The collection of these clips was carried out by requesting AMT (Amazon Mechanical Turk) workers. They were guided to pick short snippets depicting single activity and mute the audio. Each video clip duration is 10 to 25 seconds on average. Afterward, these snippets were labeled with multi-lingual, mono-sentence captions provided by the annotators. On average, it took annotators 80 seconds to complete the task, including the time required to watch the video. Precisely, there exist approximately 40 English descriptions per video snippet. Dataset has an incredible vocabulary of 16k exclusive words with eight words per sentence on average. Frequently used slices of the dataset for training, validation, and testing are 1200, 100, and 670 video clips, respectively. Some illustrative sample snippets from MSVD dataset with their available description is shown in Figure-3.

B. MSR-VTT (MICROSOFT RESEARCH-VIDEO TO TEXT)

MSR-VTT dataset [57] is an open domain, large scale benchmark with 20 broad categories and diverse video content for bridging vision and language. It comprises 10,000 clips

TABLE 1. Benchmark Datasets.

Domain/Category	Dataset	S.S	A.L(s)	T.L(hr)	Videos	Clips	S	Words	Vocab
Cooking	MPII Cooking [47]	AMT	600	8	44	-	5609	-	-
	YouCook [48]	AMT	-	2.3	88	-	2668	42,457	2,711
	TACoS [49]	AMT	360	15.9	127	7206	18,227	146,771	28,292
	TACos-MultiLevel [50]	AMT	360	27.1	185	14,105	52,593	2,000	-
	YouCook II [51]	-	316	176	2k	15.4k	15.4k	-	2,600
Movie	MPII-MD [52]	DVS, Script	4	73.6	94	68,337	68,375	653,467	24,549
	M-VAD [53]	DVS	6	84.6	92	48,986	55,904	519,933	18,269
Social Media	ActivityNet-Entities [54]	-	180	-	14,281	52k	-	-	-
	VideoStory [55]	-	-	396	20k	123k	123k	-	-
Human	Charades [56]	AMT	30	82.01	9,848	-	27,847	-	-
Multi Category	MSVD [46]	AMT	10	5.3	-	1970	70,028	607,339	13,010
	MSRVTT [57]	AMT	20	41.2	7,180	10k	200k	1,856k	29,316
	VTW [58]	Editor	90	213.2	18,100	-	44,613	-	-
	ActivityNet Captions [36]	AMT	180	849	20k	-	100k	1,348k	-
	VATEX [59]	AMT	15(E)	-	41,269	-	41,2690(E)	-	-
			13(C)				41,2690(C)		
E-Commerce	BFVD [60]	-	12*	140.4	43,166	-	43,166	-	30,846
	FFVD [60]	-	28*	252.2	32,763	-	32,763	-	34,046
Gen+Cooking	ViTT [61]	YT8M+Ann	-	-	8,169	88.5k	-	56,027ˆ	12,509
TV Shows	TVC [62]	AMT	-	-	-	108k	262k	-	-

S.S: Sentence Source, A.L: Average Length(sec), T.L: Total Length(hr), S: Number of Sentences,(E): English, (C): Chinese, *: Manually Calculated
YT8M+Ann:Videos Sampled from YouTube8M dataset and sentences from annotators, ˆ: Unique Tags

which are originated from 7180 videos. Being open-domain includes video from categories like music, people, gaming, sports, news, education, vehicles, beauty, and advertisement. The duration of each clip, on average, is 10 to 30 seconds resulting in a total duration of 41.2 hours. To provide good semantics of a clip, 1327 AMT workers were engaged to annotate each clip with 20 natural sentences. There are 200K clip-sentence pairs with 1.8M total words and 29316 distinctive words. Data split in [57] suggests 65% (6513 videos) for training, 5% (497 videos) for validation and 30% (2990 videos) for testing purposes. Figure-4 represents an example video with provided reference caption from the MSR-VTT dataset.

C. VTW (VIDEO TITLES IN THE WILD)

VTW dataset [58] comprises 18100 automatically crawled user-generated videos (UGVs) and titles. The average duration of each video clip is 90 seconds, and a single description is provided for every clip. In order to encourage the generation of diverse captions and learn sentence structure for title generation, the sentence augmentation method is introduced, which describes information not presented through the video's visual content. The dataset also provides additional description sentences with comprehensive information about each video, along with the title generation.

D. ActivityNet-CAPTIONS

ActivityNet-Captions [36] is dataset specific to dense captioning events. It covers a wide range of categories. It comprises 20k videos taken from the activity net, centered around human activities with a total duration of 849 hours and 100k descriptions. The 112 AMT workers annotated the videos. Overlapping events occurring in the video are catered, and each description uniquely describes a dedicated segment of the video, so describe events that span over time. On average, each description is made up of 13.48 words and approximately covers 36 seconds of the video. Temporally localized descriptions are used to annotate each video. On average, each video is annotated with 3.65 sentences and 40 words. Event detection is demonstrated in small clips as well as in long video sequences.

E. MP-II (MAX PLANK INSTITUTE FOR INFORMATICS)

MP-II Cooking dataset [47] distinguishes 65 fine-grained cooking activities with low inter-class and high intra-class variability, continuously recorded in a realistic setting by 12 participants preparing 14 different dishes. Activities such as pour, spice, cut slices are included. Forty-four videos are recorded with a total length of more than eight hours, and the average duration per clip is approximately 600 seconds. Activities annotation was performed by six persons resulting in 5609 annotations for all 65 activity categories.

F. YouCook

YouCook, the dataset [48] comprises 88 YouTube cooking videos, evenly split into six different cooking styles. All videos are from a third-person camera view, as shown in Figure-7, whereas MP-II cooking dataset videos are

1. A band is performing in front of a live audience.
2. A band is playing concert in open space with a huge audience.
3. There is a band playing i love rock and roll on stage.
4. People play music on a stage in front of a large crowd.
5. Some people are performing on stage.

FIGURE 4. Example video frames and captions from MSR-VTT dataset.

recorded with the fixed camera installed on the ceiling of the kitchen. The 31% objects of the YouCook dataset belong to the utensils category, 38% to bowls, remaining to food, and others. AMT annotated each video with at least three sentences with a minimum of 15 words per annotation. The average number of words per description is 67. On average, there are ten words per sentence, including stop words. The average number of descriptions per video is eight. Data split is 49 videos for training and 39 videos for test purposes, respectively.

G. YouCook II

YouCook II, [51] the dataset comprises of 2k YouTube videos that are almost uniformly distributed over 89 recipes from major cuisines of Africa, America, Asia, and Europe, having a wide variety of cooking styles, ingredients, recipes, and utensils. Each video in the dataset contains 3 to 16 temporally localized segments. These segments are annotated in English. There are 7.7 segments per video on average. The duration of each video is 5.27 minutes, with 175.6 hours of dataset length. Segments length falls between 1 to 264 seconds. About 2600 words are used while describing the recipes. Data split is 67% videos for training, 23% for validation, and 10% for testing purposes.

H. TACoS (TEXTUALLY ANNOTATED COOKING SCENES)

TACoS [49] It is a dataset derived from MP-II Cooking Composite Activities dataset [63]. Videos recording setup and annotation procedures are the same as the MP-II cooking dataset [47]. Videos are ranging from 1 minute to 23 minutes, with an average length of 4.5 minutes per video. TACoS dataset is created by filtering through MSR-VTT dataset [57]. It contains 127 videos with 26 fine-grained cooking-related activities. AMT workers provided the alignment of videos with their corresponding sentences. Twenty diverse textual descriptions were collected for each video. A total of 146771 words are used, forming 11796 sentences. Start and stop timestamps are used to align the description of the activities.

I. TACoS-MultiLevel

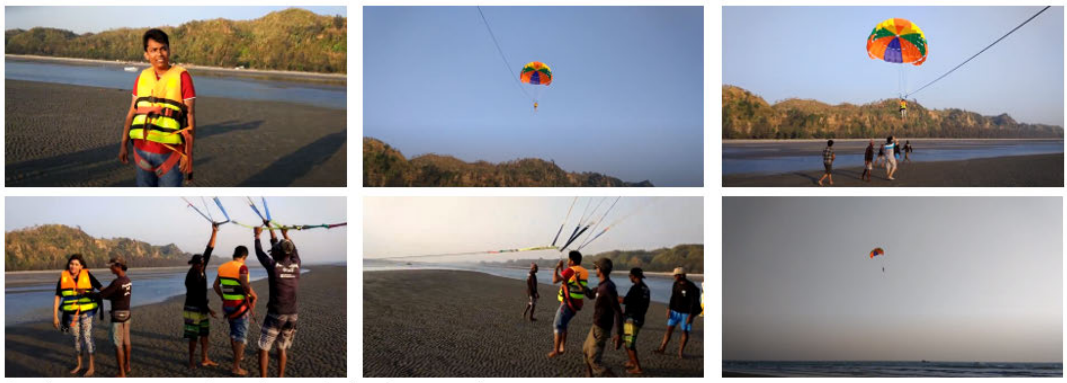
TACoS-MultiLevel [50] provides a description in a coherent way and at a varying level of detail. Text descriptions of the videos in the TACoS corpus were collected via AMT worker. Workers were guided to describe the video in detail with at most 15 sentences, describe the video in short with 3 to 5 sentences, and describe the video in a single sentence. Dataset has about 20 triples of descriptions for each video.

J. MPII-MD (MAX PLANK INSTITUTE FOR INFORMATICS - MOVIE DESCRIPTION)

MPII-MD dataset [52] is a recent collection of transcribed Audio Descriptions (ADs) consisting of 68K sentences and video clips from 94 HD Hollywood movies. Further, these movies are divided into 68337 video clips. On average, each clip is of 3.9 seconds duration, and the total duration is almost 73.6 hours. The total vocabulary size is 6,53,467, with 68,375 sentences, almost one sentence for each clip (68,337 clips and 68,375 sentences). While setting up the audio and video content configuration, every sentence was manually aligned with the corresponding video clip. In order to help visually impaired persons, the audio description track is utilized for the description of the visual contents. Data split for training, validation, and testing are 56,861 videos, 4,930 videos, and 6,584 videos, respectively.

K. M-VAD (MONTREAL VIDEO ANNOTATION DATASET)

M-VAD [53] comprises of about 49K video snippets extracted from 92 different genre movies. It is based on



1. A person is parasailing above a body of water and landing on a beach.
2. Someone is recording people who are parasailing and people who are watching too.
3. A man is riding a parachute and a group of people are standing down and watching them.
4. Someone parasailing over a lake with several men watching.
5. A person is coming down from a sky riding on a balloon glide.
6. Men on a beach prepare to assist an incoming parasailor.
7. A person is landing with a parachute onto a beach while others are greeting him or her.
8. Someone hanging from a parachute is being pulled on a line while people watch.
9. Tied to the end of a long cable, someone is para sailing and comes for a landing on a sandy beach in front of others.
10. A group of people help a person parasailing to the ground.

1. 一群人看另一个人从降落伞上准备落下。
2. 一群人看着一个人带着降落伞从空中落了下来。
3. 一个女人在一个滑翔伞上滑翔，几个男的把她拽了下来。
4. 一个人乘着降落伞即将降落到沙滩上，沙滩上的人们在对他挥手。
5. 在一个晴朗的天气，有一个人飘在空中，旁边有一些人在看着。
6. 一群人看另一个人从降落伞上准备落下。
7. 一群人看着一个人带着降落伞从空中落了下来。
8. 一个女人在一个滑翔伞上滑翔，几个男的把她拽了下来。
9. 一个人乘着降落伞即将降落到沙滩上，沙滩上的人们在对他挥手。
10. 在一个晴朗的天气，有一个人飘在空中，旁边有一些人在看着。

FIGURE 5. Example video frames and captions (English + Chinese) from VATEX dataset.

Descriptive Video Service (DVS) encoded DVDs. Its narrations are an appealing source of data for the video-sentence large paired dataset. The average duration of each video clip is 6.2 seconds, with a total duration of 84.6 hours. A total of 55,904 sentences are provided for 49k video snippets, in which some videos have more than one sentence. Data split for training, validation, and testing are 38,949 videos, 4,888 videos, and 5,149 videos, respectively.

L. ActivityNet ENTITIES (ANet-ENTITIES)

ActivityNet Entities, The large scale ActivityNet-Captions [36] dataset comprises of 20k videos from ActivityNet [64] but lacks grounding annotations; therefore, bounding box annotations are created at the entity level. 15k videos with 158k bounding box annotations from the ANet-Entities dataset. The description is accompanied by grounding and region attention. Dataset evaluates how well the generated captions are grounded. 15k videos are distributed as 10k videos for training, 2.5k videos for validation and testing each. Description quality and grounding accuracy are key characteristics of this dataset.

M. VideoStory

VideoStory [55] it is a collection of videos posted publicly on social media with diverse topics, variable lengths, high quality, and multiple viewpoints. In total, the dataset consists of 20k videos with a duration ranging from 20 to 180 seconds and provides a paragraph or multi-sentence description. Each sentence in the paragraph is aligned with the timestamps in the video. Each paragraph has 4.67 temporally localized sentences on average. Dataset has a total of 26,245 paragraphs with 123k sentences. On average, each sentence has 13.32 words. Each video has an average paragraph length of 62.23 words. Each sentence is aligned to a clip of an average of 18.33 seconds, covering 26.04% of the full video on average. Simultaneous or co-occurring events cause 22% of temporal description overlap. Data split is 17,098 videos for training, 999 videos for validation, 1,011 videos for testing, and 1,039 videos for blind-test, respectively.

N. CHARADES: COLLECTION OF CASUAL DAILY ACTIVITIES

Charades, known as Hollywood in Homes approach [56], comprising of 9848 annotated videos recorded in 15 different indoor scenes with an average duration of 30 seconds

TABLE 2. Training, Validation and Test Split Size of Benchmark Datasets.

Dataset	Train	Validation	Test
MPII Cooking [47]	1017*	-	1277*
YouCook [48]	49	-	39
TACoS [49]	-	-	-
TACos-MultiLevel [50]	-	-	-
YouCook II [51]	1340	460	200
MPII-MD [52]	83	4	7
M-VAD [53]	39k	4.9k	5k
ActivityNet-Entities [54]	7,140	3,570	3,570
VideoStory [55]	17,908	999	1011
Charades [56]	7,879	-	1,969
MSVD [46]	1,200	100	670
MSRVTT [57]	6,513	497	2,990
VTW [58]	14,480	1,810	1,810
ActivityNet Captions [36]	10k	5k	5k
VATEX [59]	25,991	3,000	12,275
BFVD [60]	28,058	2,158	12,950
FFVD [60]	21,554	1,658	9,948
ViTT [61]	5,840	1,102^	1,094^

* : Pose Estimation frames, All values represent number of videos.

^://github.com/google-research-datasets/Video-Timeline-Tags-ViTT

for each video. A total of 267 people from 3 continents contributed to the creation of this dataset. Charades provides 27,847 video descriptions, 66,500 temporally localized intervals for 157 action classes and 41,104 labels for 46 object classes. Data split is 7,985 videos for training and 1,863 videos for testing purposes.

O. VATEX (VIDEO AND TEXT)

VATEX [59] is a multilingual, large, complex, and diverse dataset for video description. It contains over 41,269 unique videos covering 600 human activities reused from a widely used benchmark for action classification dataset, kinetic-600 [65]. Kinetic 600 consists of 600 human activities comprising of 500k video clips. The average length of each clip is around 10 seconds, taken from a unique video. There exist 10 English and 10 Chinese captions with at least ten words for English and 15 words for Chinese caption for every clip in the dataset. VATEX comprises 413k English and 413k Chinese captions with 41.3k unique videos from diverse 600 human activities. A 2,159 qualified AMT English-speaking workers annotated 4,12,690 valid English captions. 450 Chinese workers write 4,12,690 valid Chinese captions. Chinese descriptions for each video clip are divided into two parts; half of the descriptions directly describing the video content while the other half is the paired translation (translation done through Google, Microsoft, self-developed translation system) of English description of the same clip. Since half of the Chinese captions are paired translations of the English captions, the total translation pairs are 2,06,345. The Figure-5 expresses the concept.

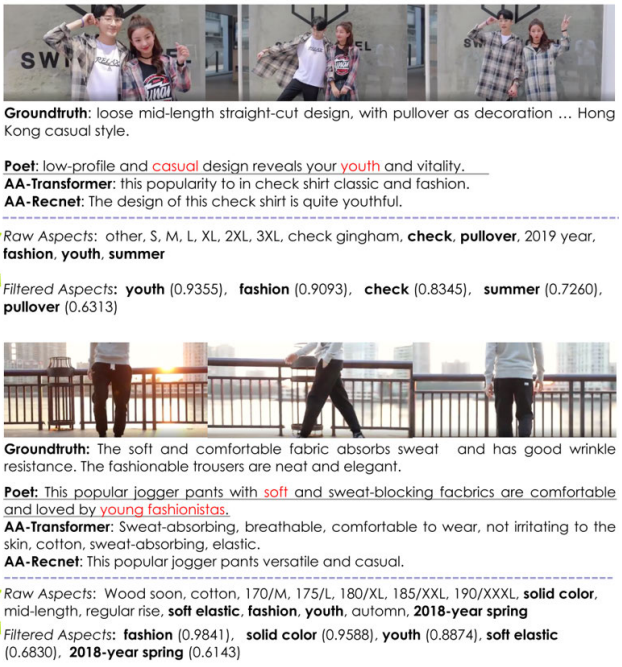


FIGURE 6. Samples generation on the FFVD test dataset [60].

P. BFVD & FFVD (BUYER-GENERATED FASHION VIDEO DATASET & FAN-GENERATED FASHION VIDEO DATASET)

BFVD & FFVD, Large scale product-oriented video captioning datasets proposed by POET [60], for video captioning in the field of e-commerce. The videos having page views over 1,00,000 and a click-through rate of more than 5% are collected from mobile Taobao (a Chinese shopping website) and labeled as either buyer or fan-generated. There are 43,166 videos of 140.4 hours duration and 32,763 videos of 252.2 hours in BFVD and FFVD. A considerable number of unique words in the datasets make it among the largest datasets for the task. Figure 6 represents the POET's [60] sample generation on FFVD test dataset in comparison to AA-Transformer [60] and AA-Recnet [60] models.

Q. ViTT (VIDEO TIMELINE TAGS)

Aiming to fix the uniformly distributed nature of YouCook II videos, the Video Timeline Tags (ViTT) dataset [61] is introduced by sampling instructional videos, particularly with cooking/ recipe labels from YouTube-8M [66] dataset. The Dataset contains 8,169 videos, of which 3,381 are related to the cooking domain. On average, there are 7.1 segments per video. 20% of captions are single-word, 22% are double-word, and 25% are three-words. There are 56,027 unique tags with a vocabulary size of 12,509 token types over 88,455 segments. After video identification, timeline annotations and descriptive tags are collected. Each step of the instructional video was identified by the annotators and assigned a descriptive yet concise tag.

TABLE 3. TVC [62] dataset split details.

Split	Videos	Clips	Desc	Desc/Clip
Train	17,435	86,603	174,350	2
Validation	2,179	10,481	43,580	4
Test-Public	1,089	5,420	21,780	4
Test-Private	1,090	5,422	21,800	4
Total	21,793	107,926	261,510	-

Desc: Descriptions

R. TVC (TV SHOW CAPTION)

TV Show Caption dataset [62] is a multimodal captioning dataset with 262K captions created by extending the TVR (TV show Retrieval) dataset by storing additional descriptions for every single annotated video clip or moment. Similar to the TVR dataset, the TVC task involves utilizing both video and subtitles for required information collection and appropriate descriptions generation. The TVC contains 108K video clips paired with 262K descriptions, and on average, there exist two to four descriptions per video clip. Since TVC is created on top of the TVR, it relates in many ways with the TVR, like a variety of actions and people in a single description, language range and diversity, and intense inter-human connections and interactions. The human annotators were engaged in writing descriptions for video only and video+subtitle if a subtitle already exists. The TVC description type distribution shows a balance compared to TVR with 50% descriptions belonging to the video only, whereas around 31% belong to both video and subtitles, and 18% of descriptions only come from subtitles. The statistics for train/validation/test-public/test-private split can be viewed in Table-3. The transformer-based MMT model [62] evaluated on TVC with both video and subtitles modalities outperformed the models with one of these modalities. It established the fact that both videos and subtitles are equally valuable for concise and appropriate description generation. Unlike previous datasets employed for video description focusing on captions illustrating the visual content, the TVC dataset aims at captions that also describe subtitles.

IV. DISCUSSION ON DATASETS

Comparable to countless emerging deep learning-influenced fields, visuolingusitic systems' success undoubtedly counts on the training data both in terms of quantity and quality. Currently, available datasets and the systems evaluated using these datasets are in their formative or infancy phase. Due to the uncomplicated and straightforward nature of the videos and their reference captions, it is a bit difficult to produce coherent and concise captions. Increasing the difficulty levels of the task-oriented and standardized datasets along with the advancement in annotation practices employing rich text aligned with the videos fulfilling the primary attributes of reliability, genuineness, and diversity can enhance the system performance. A few parameterized datasets are designed explicitly in dense video captioning domain for generating more natural, coherent, and diverse descriptions, like ActivityNet Captions, YouCook-II, and VideoStory. As a next step to improve the generated captions, the combination of subtitles, if available, with the visuals also contribute to improving the videos' information-seeking capabilities like the large-scale TVC and diverse ViTT dataset establishing new benchmarks for driving more progress in this direction. Likewise, the accessibility of acoustic attributes of videos can also contribute to performance enrichment. Although researchers have efficiently processed and conducted the field's technicalities and versatile models have been proposed for captioning employing convolutions, recurrence, attention mechanism, reinforcement learning, transformers approach with the above-listed datasets. Furthermore, the inconsistencies and discrepancies between the generated and actual human-annotated descriptions diminish with a consistent stride. However, to produce further human-like explanations and make the systems significantly effective, sizeable task-orientated and standardized datasets exploiting text, i.e., subtitles and audio already present within the videos and extending substantial ground truth references for effective training/validation/testing practice is indispensable.

V. EVALUATION METRICS

Video description is a joint venture of CV and NLP, whereas the metrics commonly used for the evaluation of automatically generated captions, i.e., BLEU, METEOR, ROUGE, and WMD, are from the NLP domain, i.e., MT related, and document summarization. Machine translation systems, evaluated using these automatic metrics, produce faster, easier and cheaper evaluations as compared to the evaluation performed by trained, bilingual human evaluators who can accurately assess the predicted translations [67]. During the rise of the image captioning task, the need arise for task-specific evaluation metrics, which can gauge the performance of the designed model. As a result, CIDEr [68] and SPICE [69] evaluation metrics evolved. All these metrics are considered for the evaluation of video descriptions in the literature. For the video to be described using deep learning techniques, single or multiple annotations or ground truth sentences per video clip are available for the purpose of comparison with the generated description. These provided human annotations work as a reference while evaluating predicted descriptions. Adequacy, fidelity, and eloquence of the translation are the main aspects of machine translation observed by humans to do the evaluation [70]. The most desirable characteristic of an automatic evaluation metric is its strong correlation with human scores [71], i.e., the closer the generated or predicted translation to a professional human translation is considered better. The accuracy of a metric is considered to be higher if it assigns a greater score to the caption favored by humans [70]. A brief description and computation concept, along with limitations exhibited by these automatic evaluation metrics, are also summarized in table-4. Each metric in detail is given below:

1. A lady pours something like oil into a bowl along with some other liquid and mixes it well. She drops beans and mint inside along with some other ingredient which is in liquid form. She mixes it well and adds cheese slices to it and mixes it again.
2. A lady is pouring oil in a small glass bowl. After that she is pouring jelly like thing in that. Then honey like thing in it. She is mixing it well with a spoon. She is putting green peas in a big glass bowl, putting green leaves in top of that. Pouring already blended oil in that big bowl and mixing it well.
3. We start in a woman's kitchen. She adds oil to a bowl and some lemon juice. To that she adds some honey and mixes it all together. We see a note that pops up to tell us that Edaname is rich in carbohydrates, proteins, fiber and fatty acids. Once it is mixed together, she adds the beans to a bowl along with some mint, salt, pepper, and the dressing that she previously made. this is all stirred together and topped with some cheese slivers that get mixed into the salad.
4. In this video there is a kitchen background and a women mixing vegetables and oils. She first pours some kind of oil into the bowl and explains what to do next. Then she adds a type bean or vegetable into the bowl and keeps stirring. Then she is finished making her mixture.
5. In this video, a woman pours liquids form three small bowls into a medium sized bowl then mixes it well and keeps it aside. Then in large bowl she adds all other ingredients like some green color seeds, leaves and mixes it. Then she adds mixed liquid to it. An oven, frying pans and fruits in a plate and flowerpot can be seen in the background.

FIGURE 7. Example video frames and captions from YouCook dataset.

TABLE 4. Evaluation Metrics.

Metric	Domain	Computation Concept	Limitation
BLEU [72]	Machine Translation	n-gram precision	Lack of Recall, Weakly co-related with the human judgment of translation quality [28]
METEOR [67]	Machine Translation	n-gram with synonym matching	Sensitive to n-gram overlap [69]
ROUGE-L [73]	Document summarization	n-gram recall	Weakly co-related with human judgment, All n-grams have to be continuous [74]
WMD [75]	Document Similarity	Earth Mover's Distance (EMD) on word2vec	-
CIDEr [68]	Image description generation	tf-idf* weighted n-gram similarity	unbalanced tf-idf* weighting [76]
SPICE [69]	Image description generation	Scene-graph synonym matching	fails to capture the syntactic structure of a sentence [28]
SODA [77]	Dense Video Captioning	F-measure using precision & recall	-

*tf-idf: term frequency-inverse document summary [78]

A. BLEU (BILINGUAL EVALUATION UNDERSTUDY)

This well-known evaluation metric was proposed by [72]. The central concept behind the renowned evaluation metric BLEU is the measurement of the numerical closeness of generated translation with the provided reference annotation. BLEU computes the adequate overlap of a single word, i.e., unigram or adjacent multiple words, i.e., n-gram between the automatically generated caption and the provided reference human-annotation. Alternatively, we can say it is defined as the geometric mean of the n-gram match

count. A variant of the BLEU metric, referred to as the ''NIST'' metric, was proposed by [79]. BLEU was particularly designed to evaluate short sentences, so evaluating complex or multi-sentence captioning using BLEU makes it difficult to evaluate accurately. Providing a single video with more reference human annotations will increase the probability of getting a good BLEU score. The cornerstone of this metric is the precision measure. Predicted translations shorter than the ground truth references are penalized by modified n-gram precision. BLEU uses *Brevity Penalty* (*BP*), which penalizes generated translations for being too short. *BP* can be computed as (1).

$$BP(P,R) = \begin{cases} 1, & ifLP > LR \\ e^{(1-\frac{LP}{LR})}, & ifLP \leq LR \end{cases} \quad (1)$$

where *LP* is the length of the predicted translation and *LR* is the length of the reference annotation. The Overall BLEU is calculated using the geometric mean of the n-gram precision (p_n) as shown in (2) and (3).

$$logBLEU = min\left(1 - \frac{LP}{LR}, 0\right) + \sum_{n=1}^{N} (w_n log p_n) \quad (2)$$

$$logBLEU = BP + ActualMatchScore \quad (3)$$

where w_n represents positive weights summing to 1 using n-grams up to length N in (2). One major limitation of BLEU is the only consideration of precision and lack of recall measures. [80] showed that a significantly better correlation could be obtained by emphasizing more on recall than on precision.

B. METEOR (METRIC FOR EVALUATION OF TRANSLATION WITH EXPLICIT ORDERING)

Meteor evaluation metric was proposed by [67]. The basis for this metric lies in an explicit exact word matching between the predicted translation and the single or multiple reference annotations. Matching of words supports identical words and the words with the identical stem as well as synonyms. The score is calculated by comparing the generated sentence with the best matching among all the reference sentences. The computation idea behind this popular metric is the harmonic mean of precision and recall of uni-gram matches between sentences [76]. The limitation of BLEU, lack of recall, is tried to be compensated by this metric. It captures the matching as well as the order of the words in predicted and reference sentences. WordNet (WN) [81], a lexical database for English, acts as a language resource, containing information about around 155,000 nouns, verbs, adjectives, and adverbs, including simple words, phrasal verbs, and idioms are used for matching purposes have dramatically improved the evaluation accuracy [82]. Alignment of predicted and reference sentences are performed in three stages, i.e., exact mapping, porter mapping, and WN-stem mapping while calculating the score.

METEOR score for predicted and reference sentences is calculated using unigram precsion P as in (4) and unigram recall R as in (5) where precision is the ratio of the number of unigram co-occuring in both predicted and reference sentence UG_{PR} to the number of unigram in the predicted sentence UG_P.i.e

$$Precision = P = \frac{UG_{PR}}{UG_P} \quad (4)$$

The recall is the proportion of the unigram co-occurring in both predicted and reference sentence UG_{PR} to the number of a unigram in the reference sentence UG_R, i.e.

$$Recall = R = \frac{UG_{PR}}{UG_R} \quad (5)$$

where UG_{PR} represents number of unigram co-occuring in predicted and reference sentences, UG_R represents number of unigram in reference sentences and UG_P represents number of unigram in predicted sentences. Mean harmonic score (F), using precision and recall is calculated as (6):

$$F_{mean} = \frac{10PR}{R + 9P} \quad (6)$$

The Penalty is computed and applied when taking longer matches and non-adjacent mappings between the predicted and reference sentences. Unigrams in predicted sentence that are mapped to the unigrams in reference sentences are grouped into chunks, resultantly longer n-gram will have fewer chunks. Penalty (P_n) can be calculated as (7):

$$P_n = 0.5 * \frac{C}{UM} \quad (7)$$

where C represents the number of chunks and UM represents the number of unigrams matched. Finally the METEOR score for given alignment can be computed as (8):

$$METEOR = F_{mean}(1 - P_n) \quad (8)$$

It reduces the F_{mean} by a maximum of 50% if no longer matches are there. Using *spearsman's correlation analysis*, [83] showed that METEOR is supposed to be more intensely correlated with human judgment than the BLEU score and the Meteor score is higher the better.

C. ROUGE (RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION)

Rouge evaluation metric [73] initially designed for evaluation of document summaries. Summaries evaluated by humans require semantical coherence, conciseness, grammatical correction, readability, avoiding redundancy, and, most importantly, content itself as well as the logical organization of the content [74] but consumes time. After the application of automatic evaluation metrics to Machine Translation, it was shown by [84] that the same might be applied to document summarization. Rouge is a package containing multiple variants used to measure the similarities between the generated and reference summaries. These variants are Rouge-N

(n-gram Co-occurrence), Rouge-L (Longest Common Subsequence), Rouge-W (Weighted Longest Common Subsequence), and Rouge-S (Skip-bigram Co-occurrence). We will review in detail only Rouge-N and Rouge-L, as they help evaluate the image and video caption.

1) Rouge-N: It can be defined as an n-gram recall between the predicted summary and one or multiple reference summaries. It can be calculated as (9):

$$R_N = \frac{\sum_{R \in RS} \sum_{g \in R} C_M(gr_n)}{\sum_{R \in RS} \sum_{g \in R} C(gr_n)} \tag{9}$$

where RS refers to reference summaries, n stands for length of n-gram, gr_n, and $C_M(gr_n)$ is the maximum number of n-gram co-occurring in a predicted summary and collection of reference summaries. As we add more summaries in the set of reference summaries, the number of the n-gram in the above formula's denominator also increases.

2) Rouge-L: It is the variant of Rouge can be used to evaluate image and video captioning. It uses recall and precision value of the *longest common sub-sequence* between the generated and each reference sentence. The perception is that longer LCS of predicted and reference sentences will generate high similarity score. *Recall* computed in (10), *precision* computed in (11) and *F-measure* score in (12) on the basis of LCS for predicted summary P_s of lengthL_p and reference summary R_s of length L_r can be computed as:

$$R_{LCS} = \frac{LCS(R_s, P_s)}{L_r} \tag{10}$$

$$P_{LCS} = \frac{LCS(R_s, P_s)}{L_p} \tag{11}$$

$$Rouge_{LCS(R_s,P_s)} = F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{12}$$

where $LCS(R_s, P_s)$ represents the longest common subsequence of R_s and P_s. β is the ratio of *LCS-Precision* to *LCS-Recall* i.e. $\beta = \frac{P_{LCS}}{R_{LCS}}$. We can see that $LCS(R_s, P_s) = 1$ when both predicted and reference summaries are same and $LCS(R_s, P_s) = 0$ when there is nothing similar between the two summaries.

D. WMD (WORD MOVER'S DISTANCE)

WMD proposed by [75] represents distance function between text documents. WMD distance is a measure of dissimilarity between two text documents. To resolve two sentences with all different words that may convey the same meaning, likewise two sentences having the same objects, object-relationship, and characteristics but conveying different meanings, WMD was proposed. The metric leverages results for computing the transportation cost from word2vec [85] embedding. Text documents are presented as embedded word's weighted point cloud. WMD provides substantial advantages over other metrics [76].

E. CIDEr (CONSENSUS-BASED IMAGE DESCRIPTION EVALUATION)

CIDEr is an evaluation protocol proposed to evaluate image description that uses human consensus [69]. The core idea behind this evaluation metric is the measure of similarity of a predicted sentence against a single or a set of reference captions provided for the image by human annotators. Therefore, in sentence similarity, the concepts of prominence, accuracy, grammatical, and significance are intrinsically apprehended by the CIDEr. It is an extension of tf-idf weighing mechanism from information retrieval where common n-grams in all image captions are penalized. As cosine similarity is used to compute CIDEr, therefore, at times, some insignificant but repeatedly used fragments of the captions get extraordinary weightage resulting in effective evaluation. To show high agreement with accurate consensus, captions per image as references need to be in high number, i.e., dataset ABSTTRACT-50S with 50 reference captions per image was used by [68], based on the dataset of [86]. All words in both predicted and reference sentences are mapped to their stem/basic/root form, i.e., 'walks', 'walked', and 'walking' is mapped to 'walk'.

For evaluation of an image I_i automatically, the predicted description P_i matches the consensus of a set of reference sentences S_i such that $S_i = \{S_{i1}, S_{i2}, \ldots S_{im}\}$. $Cider_n$, score for n-gram of length n can be computed as average cosine similarity between predicted and reference captions as in (13).

$$CIDEr_n(p_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(p_i).g^n(S_{ij})}{\|g^n(p_i)\| \|g^n(S_{ij})\|} \tag{13}$$

where m is the number of reference captions, $g^n(p_i)$ is a vector representing all n-gram with length n and $\|g^n(p_i)\|$ is the magnitude of $g^n(p_i)$, similar is the case for $g^n(S_ij)$. Using these parameters $CIDEr$ score can be computed as (14):

$$CIDEr_n(p_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(p_i, S_i) \tag{14}$$

where it was observed by the [68] that uniform weights $w_n = \frac{1}{N}$ and $N = 4$ works best.

A variant of CIDEr, famous for image and video description evaluation, is *CIDEr-D*. Modifications carried out in basic *CIDEr* was the removal of stemming, which ensured the usage of correct forms of words. Repetition of higher confidence words was avoided by introducing a Gaussian penalty on the basis of difference in predicted and reference captions. Finally, adding clipping to the n-gram count in $CIDEr_n$ numerator was done as a sentence length penalty. It was shown that *CIDEr-D* variant has a high *spearman's rank correlation* with respect to original *CIDEr* score.

F. SPICE (SEMANTIC PROPOSITIONAL IMAGE CAPTION EVALUATION)

SPICE evaluation metric was proposed by [69] to measure the quality of predicted or generated captions by exploring their semantic contents. For automatic evaluation, n-gram overlap

is neither essential nor adequate for both predicted and reference captions to communicate the same [87]. In order to overcome the limitations exhibited of n-gram overlap by the above-defined automatic evaluation metrics, propositional semantic content is emphasized. A graph-based semantic representation, known as a scene graph, is used for the assessment of the quality of captions. These semantic scene graphs allow for noun and object matching in the captions. BAST [88] is closely related to SPICE for captioning.

The objective of SPICE is to calculate a score that captures the similarity between predicted captions and a set of reference captions S where $S = \{S_1, S_2, \ldots S_m\}$. Scene graph for predicted caption p is $G(p)$ whereas scene graph for set of reference captions S is $G(S)$ where $G(S)$ is union of all $G(Si)$. A scene graph tuple $G(p)$ shown in (15) of predicted caption p consists of semantic components such as object classes $O(p)$, relation types $R(p)$ and attribute types $A(p)$, parsing c to scene graph:

$$G(p) = \langle O(p), R(p), A(p) \rangle \tag{15}$$

A function T is defined to get logical tuples from scene graph and $\otimes$ acts as binary matching operator, then precision P, recall R and $SPICE$ score can be computed as (16), (17), and (18):

$$Precision = P(p, S) = \frac{|T(G(p)) \otimes T(G(S))|}{|T(G(p))|} \tag{16}$$

$$Recall = R(p, S) = \frac{|T(G(p)) \otimes T(G(S))|}{|T(G(S))|} \tag{17}$$

$$SPICE = F_1(p, S) = \frac{2.P(p, S).R(p, S)}{P(p, S) + R(p, S)} \tag{18}$$

Fluency adjustment is not included while computing the $SPICE$ score because it is assumed to be a conceptually simple and easily interpretable metric. It uses WordNet [81] synonym matching approach similar to *Meteor* [67] metric.

G. SODA (STORY ORIENTED DENSE VIDEO CAPTIONING EVALUATION FRAMEWORK)

The current dense video captioning evaluation framework, an extension of dense image captioning (DIC), is insufficient for evaluating video story descriptions. DIC does not consider the temporal dependency between captions explicitly, which causes the potential risk of overestimation. Another problem associated with the current evaluation framework is assigning a high score for producing a high number of inadequate captions. Loose matching,i.e., matching a generated caption with many reference ground truths or matching a reference caption with many generated captions, also adds to the inefficiency. Overestimation of METEOR is due to loose matching. Currently, averaging METEOR score cannot tackle the coverage of generated captions (recall) and the accuracy of the captions (precision). Generation of too many or too few captions makes the evaluation system inadequate. For an appropriate and correct evaluation of the video story systems, a framework is required to consider a video story, the ordering of captions, and penalize redundant captions. Dense Video Captioning (DVC) evaluation involves two significant attributes, i.e., the accuracy of localized events and the accuracy of generated captions for each event. Since the order of the events is also an essential factor, so SODA [77] believes in the ordering of captions while measuring the system performance. SODA gives low scores against too many or too few captions and high scores against captions whose number equals ground truth references. Furthermore, SODA tends to give lower scores than the current evaluation framework in evaluating captions in incorrect order.

SODA is helpful not only for the ActivityNet Captions dataset but also for the other datasets created to evaluate system captions that convey the story. It would be more tricky to obtain a factitiously high score with SODA than the current evaluation framework because SODA requires systems to detect the exact number of events and captions that lead to further progress of DVC tasks. SODA finds the best sequence of generated proposals that maximizes the sum of the IoU against reference proposals by first applying dynamic programming that finds the optimal matching between generated and reference captions considering the temporal ordering of captions. It computes METEOR scores for the matched pairs and derives precision and recall scores based on the calculated METEOR scores. The proposed framework evaluates generated captions with F-measure scores from the METEOR scores to penalize redundant captions and to consider both the numbers of generated and reference captions.

Let G be a set of manually-generated reference captions for a video and P be a set of captions generated by a system. We denote g as a reference caption and p as a caption generated by the system. Each caption has a proposal that indicates a time span of an event that appears in a video. Here, the IoU between g and p is defined as (19):

$$IoU(g, p) = max\left(0, \frac{min(e(g), e(p)) - max(s(g), s(p))}{max(e(g), e(p)) - min(s(g), s(p))}\right) \tag{19}$$

where function $s(.)$ represents the start time of the event proposal and $e(.)$ indicates the event proposal end time.

Let τ represents the IoU threshold then the set of ground truth captions with IoU exceeding τ ($IoU(g, p) \geq \tau$) can be defined as (20):

$$G_{p,\tau} = \{g \varepsilon G | IoU(g, p) \geq \tau\} \tag{20}$$

Set of generated captions, P, is evaluated based on set of reference captions, G, using the following equation (21):

$$E(G, p, \tau) = \frac{\sum_{p \in P} \sum_{g \in G_{p,\tau}} f(g, p)}{\sum_{p \in P} |G_{p,\tau}|} \tag{21}$$

where $f(g, p)$ represents an evaluation metric; METEOR is considered here.

A cost $C_{i,j}$ is defined between a reference caption g_i and a generated caption p_i based on IOU as (22):

$$C_{i,j} = \begin{cases} IoU(g_i, p_i), & if IoU(g, p) \geq \tau \\ 0, & Otherwise \end{cases} \tag{22}$$

$S[i][j]$ is defined for holding the maximum score of optimal matching between 1st and ith generated captions and 1st and jth reference captions.

The f-measure for the set of references and generated captions can be computed using Precision (as in (24)) and Recall (as in (25)) by the following formula given in (23):

$$F - measure(G, P) = \frac{2 \times Precision(G, P) \times Recall(G, P)}{Precision(G, P) + Recall(G, P)} \tag{23}$$

where

$$Precision(G, P) = \frac{\sum_{g \in G} f(g, p_{a(g)})}{|P|} \tag{24}$$

and

$$Recall(G, P) = \frac{\sum_{g \in G} f(g, p_{a(g)})}{|G|} \tag{25}$$

VI. DISCUSSION ON EVALUATION METRICS

Human beings can describe any video in various ways accurately and express what is happening appropriately. However, it is equally difficult for the machines to automatically generate a caption for a given video and even more challenging to evaluate that generated caption for its accuracy. A considerable barrier impeding progress in the video description domain is the lack of an appropriate evaluation mechanism. Since evaluation metrics available at present are from the machine translation, document summarization, and image captioning domain except for SODA, designed explicitly for dense video captioning evaluation. Limitations of these metrics include poor performance on words replacement with synonyms (BLEU, CIDEr), weak correlation with human judgment (BLEU, ROUGE), sensitivity to n-gram overlap (METEOR), failure to capture the syntactic sentence structure (SPICE), word order change (BLEU, ROUGE, CIDEr), and change in sentence length(BLEU, METEOR, ROUE). Considering these limitations, there is a severe demand for an evaluation metric that is closer to human judgments. Instead of manual computations, a reinforcement learning-based (or any machine learning or deep learning) mechanism can be employed to learn metrics by exploration/exploitation for generated caption evaluation. Although it has not been studied in the literature at this time, SODA is a considerable step towards the domain-specific evaluation metric goal and success in the field.

VII. PRE-TRAINING VISUOLINGUISTIC TASKS

The introduction of deep learning in vision and language domains employed a standard pipeline of CNN and RNN(LSTMs, GRU) for visual and language modeling for many years. The accomplishment of the deep-learning based models initially in the NLP domain and afterward in the cross-modal tasks of vision and language encouraged the researchers to enhance performance further. One of these performance enhancement techniques is pre-training; pre-training plays a vital role in boosting the performance of vision and language-related tasks. The concept revolves around pre-training the proposed model on sizeable unlabeled data and then fine-tuning the required downstream task employing the related labeled data. Pre-training can influence the performance of both understanding tasks (retrieval) and the generation tasks (captioning). BERT [95]; language modeling based on the transformer got attention for both performance enhancement due to parallelization (transformer mechanism employment) and pre-training approach. Motivated by the accomplishment of BERT for NLP tasks, several cross-modal pre-training models demonstrating the effectiveness of pre-training have been proposed recently, including VideoBERT [93], ActBERT [91], ViLBERT [96], CBT [94], HERO [22], BART [97], MASS [98], UniVL [90], VLM [89], ViTT [61], and ASR-Trf [93].

ViLBERT [96] investigated the extension of the pre-training idea presented in BERT for combined vision and language tasks instead of only language modeling. The ViLBERT's [96] compositional model of CNN and RNN with pre-training demonstrated surpassed results. To address the semantic alignment of the video and language (extracted from the audio of the same video), VideoBERT [93] proposed hierarchical vector quantization to get the visual tokens to generate caption and predict the next frame. CBT [94] working with S3D, proposed the sliding window mechanism for visual features extraction. VideoBERT and CBT are the first to explore the pre-training of language and vision on instructional videos. ActBERT [91] employed a pre-trained tangled transformer-based model intending to predict the action performed given the text and visual information, i.e., action classification. Recently, UniVL [90] explored understanding and generation tasks by employing two single-modal encoders, a cross encoder, and a decoder with a transformer backbone. Most of the proposed models are evaluated on the YouCook II dataset because cooking videos tend to possess high visual and language semantics temporally alignment probability for caption generation. The TransED [92] by pre-training on Auto-captions on GIF and then fine-tuning it on MSR-VTT, consistently presents better performances than TransED-without pre-training across all the evaluation metrics. The achieved performance confirms the merit of using vision-language pre-training over Auto-captions on GIF, which accelerates the downstream task of video captioning on both online and offline test split of the MSR-VTT dataset. The ViTT [61] model proposed separate-modality framework employing co-attention transformer. The ViTT's authors constructed their own ViTT dataset to address the generalization implications imposed by video's small size, uniform nature, and uncomplicated behavior in the YouCook II dataset. Table-5 compares the performance of some of the recently presented models employing pre-training. It shows that among the models evaluated on the YouCook II dataset, UniVL demonstrated high results for BLEU, METEOR, and ROUGE-L. For TVC evaluation, HERO outperformed the ActBERT for BLEU, METEOR, and CIDEr.

TABLE 5. Performance comparison of cross-modal models employing pre-training.

Ref	Year	Model	Dataset		PT	Modalities	B@3	B@4	M	R	C
			Evaluated On	Pre-Trained On							
[89]	2021	VLM	YouCook II	HowTo100M	✓	V + Tx	17.78	12.27	18.22	41.51	1.39
[90]	2020	UniVL	YouCook II	HowTo100M	✗	-	14.23	9.46	16.27	37.44	1.15
					✓	T	20.32	14.70	19.39	41.10	1.81
					✓	V	16.46	11.17	17.57	40.09	1.27
					✓	V + T	23.87	17.35	22.35	46.52	1.81
[22]	2020	HERO	TVC	HowTo100M	✓	V+Sb	-	12.35	34.16	17.64	49.98
[91]	2020	ActBERT	TVC	HowTo100M	✓	G+L+Tx	8.66	5.41	13.30	30.56	0.65
[61]	2020	ViTT	YouCook II	YT8M-cook, Recipe1M	✗	ASR+V	-	8.01	16.19	34.66	0.91
					✓	ASR+V	-	12.04	18.32	39.03	1.23
					HB	-	-	15.20	25.90	45.10	3.80
[61]	2020	ViTT	ViTT-All	HowTo100M, WikiHow	✗	ASR+V	-	19.49	9.23	28.53	0.69
					✓	ASR+V	-	22.45	11.0	31.49	0.82
					HB	-	-	43.34	33.56	41.88	1.26
[61]	2020	ViTT	ViTT-Cooking	YT8M-cook, Recipe1M	✗	ASR+V	-	24.22	12.22	32.60	0.89
					✓	ASR+V	-	24.92	12.43	33.10	0.90
					HB	-	-	41.61	32.50	41.59	1.21
[92]	2020	Trans-ED	MSR-VTT	Auto-captions on GIF	✗	V + Sen	-	38.3	26.8	59.2	44.3
					✓	V + Sen	-	39.0	27.3	59.7	45.2
[93]	2019	VideoBERT	YouCook II	Kinetics [67]	✓	VideoBERT (ASR+V)	6.80	4.04	11.01	27.50	0.49
					✓	VideoBERT (V)	6.33	3.81	10.81	27.14	0.47
					✓	VideoBERT + S3D	7.59	4.33	11.94	28.80	0.50
[93]	2019	ASR-Trf	YouCook II	-	✓	AT	-	8.55	16.93	35.54	1.06
				ImageNet	✓	AT + V	-	9.01	17.77	36.65	1.12
[94]	2019	CBT	YouCook II	HowTo100M, Kinetics	✓	V	-	5.12	12.97	30.44	0.64

T: Transcript, V: Video, PT: Pre-Training, MM: Multimodal, AT: ASR Transformer, Trf: Transformer, G: Global action features, L: Local regional features
Tx: Text, Sb: Subtitle, B@3: BLEU@3, B@4: BLEU@4, M: METEOR, R: ROUGE-L, C: CIDEr, Sen: Sentence
HB - Human Baseline is an estimate of human performance as reported by [93], and can be taken as a rough upper bound of the best performance achievable.

One of the differences between pre-training and fine-tuning datasets is the way video segments are defined. For unsupervised or pre-training datasets, the segments are defined by some empirical rule, whereas for the supervised or fine-tuning dataset, human annotators define segments corresponding to the video contents or instructional steps. Second, although both types of datasets are similar in presenting abstracts of instructional videos, they vary in style, length, and details. Despite these differences, pre-training data plays a vital role in boosting performance. A brief overview of some of the widely used video datasets for achieving the pre-training objectives is given below.

A. HowTo100M

A collection of narrated videos, particularly instructional videos containing complex tasks featuring 130M video clips [99] extracted from 1.2 million videos comprising of 12 categories of 23,611 visual tasks from YouTube associated with a manually written or ASR generated English subtitle. Videos with complex activities are relatively long, with an average duration of 6.5 minutes. Each video produces 110 clip-caption pairs with an average duration of four seconds per clip and 4 words per caption. 71% of the videos are found to be instructional.

B. Recipe1M+

A large scale, structured corpus [100] of more than one million cooking recipes along with 13 million food images scraped from popular cooking websites. The dataset content can be segregated into two groups. The first one comprises the title, list of ingredients, and steps required for certain recipes. The nutritional information is also contained in this group if the measurement of ingredients is provided. The second group contains associated images. Moreover, each recipe is categorized with a category label like an appetizer, dessert, side dish. Seventy percent of the data is labeled for the training purpose, and the rest is split uniformly between validation and test set.

C. WikiHow

A large scale summarization dataset [121] consisting of 230,843 diverse articles accompanied by their summaries

TABLE 6. Video Description - Quantitative Performance Evaluation on MSVD Dataset.

Ref	Approach/ Model	Year	B	M	R	C
[101]	AVSSN	2021	62.3	39.2	76.8	107.7
[102]	SBAT	2020	53.5	35.3	72.3	89.5
[103]	SAAT	2020	46.5	33.5	69.4	81.0
[25]	VNS-GRU	2020	64.9	**41.1**	**78.5**	**115**
[104]	JSRL-VCT	2019	52.8	36.1	71.8	87.8
[105]	GFN-POS	2019	53.9	34.9	72.1	91.0
[106]	DCM-ED (M)	2019	53.3	35.6	71.2	83.1
[26]	TDConvED(R)	2019	53.3	33.8	-	76.4
[107]	OAM	2019	43.5	31.6	-	64.9
[108]	SDN	2019	61.8	37.8	76.8	103
[24]	GRU-EVE	2019	47.9	35.0	71.5	78.1
[27]	EtENet-IRv2	2019	50.0	34.3	70.2	86.6
[109]	OA-BTG	2019	56.9	36.2	-	90.6
[110]	LR-dep(Non-Local)	2019	49.7	33.7	71.7	84.5
[111]	SibNet	2018	54.2	34.8	71.7	88.2
[112]	Tubes	2018	77.6	32.6	69.3	52.2
[113]	SeFLA	2018	**84.8**	-	-	94.3
[114]	RecNet	2018	52.3	34.1	69.8	80.3
[115]	TDDF	2017	45.8	33.3	69.7	73.0
[116]	LSTM-TSA	2017	52.8	33.5	-	74.0
[117]	S2VTK	2017	42.5	31.0	-	-
[118]	LSTM-E	2016	45.3	31.0	-	-
[119]	S2VT	2015	-	29.8	-	-
[120]	LSTM-YT	2014	33.29	29.07	-	-

(B:BLEU, M:METEOR, R:ROUGE, C:CIDEr)

extracted from WikiHow knowledge base with an average article length of 579.8 sentences and an average summary length of 62.1 sentences. The total vocabulary size is 556,461. Each article starts with the title ''How to.'' Each step starts with a step summary and is then followed by a detailed explanation of the step. ViTT model mined all step summaries comprising 1,360,145 segments with 8.2 words per segment and each of the instruction step is considered a distinct example during pre-training of the model.

D. YouTube-8M

It is the largest multi-label video classification dataset [66] consisting of around eight million videos, resulting in 500k hours of watch time, with the aim to determine the topical themes of a video. These videos are annotated automatically with the vocabulary of 4800 visual entities/classes categorized into 24 top-level classes for diversity illustration. The dataset contains frame-level features of over 1.9 billion video frames. These videos are annotated with the YouTube video annotation system to get topic annotation for a video and get a video for any given topic. Unlike typical event recognition or object detection, this dataset aims to understand what is happening in the video and afterward summarize into few key topics. A video may be annotated with more than one class. On average, there exist 1.8 classes per video. The average length of a video is 229.6 seconds.

E. KINETICS

The Kinetics dataset containing 400 human action classes with 400∼1150 video clips for each action was introduced to address the small-sized HMDB-51 and UCF-101 with insufficient variations to train and test deep learning-based human action classification models. Each video clip is from YouTube with an average duration of ten seconds having variable resolution and frame rate.

F. AUTO-CAPTIONS ON GIF

The diverse and complex video-sentence dataset [92] constructed for the video understanding task contains 163,183 GIF videos and 164,378 sentences with 31,662 vocabularies derived automatically from billions of web pages with immense video categories. The open domain nature of the dataset facilitates the generalization capability of pre-trained representation on downstream tasks. For each crawled GIF video, a corresponding sentence is selected free from polarity annotations. Sentences with a high repetition rate, absence of noun or preposition, use of high frequency but less informative phrases, and semantically mismatched GIF video-sentence pairs are filtered out. Sentences passing through human-like annotation binary classifiers, i.e., matching with human written sentences from MSVD, MSR-VTT, and MSCOCO dataset but simultaneously not matching the set of discarded sentences, are used for dataset construction.

VIII. QUANTITATIVE & QUALITATIVE RESULTS

The qualitative and quantitative results generated by various models using benchmark datasets in the recent past are discussed in this section. Segregation according to the dataset used by the model has been further categorized in chronological order accordingly. For models having multiple variants during experimentation, the best performing variant is reported here. Scores shown in bold letters are the best performing. Since all the evaluation metrics follow *the higher, the better* strategy, therefore, higher scores are considered to be better for all BLEU, METEOR, ROUGE, and CIDEr. For the models computing BLEU@1, BLEU@2, BLEU@3, and BLEU@4, only BLEU@4 is reported here because of its analogous characteristic with the human annotation. Figure 8 represents the comparison of the qualitative results of various models employing encoder-decoder architecture with a brief explanation.

MSVD and MSR-VTT datasets are famous among researchers because of their wide-ranging comprehensive categories and diverse nature of the videos and the availability of multiple ground truth captions for model training and evaluation, and most importantly, task specificity. Table-6 summarized the quantitative results of popular models on the MSVD dataset. Regarding the BLEU metric, SeFLA, where apart from visual features, semantic features are also considered, and captions are generated by employing an attention mechanism. VNS-GRU surpassed the rest of the proposed models for METEOR, ROUGE, and CIDEr scores. Similarly, Table-7 presents the scores reported using the MSR-VTT dataset where VNS-GRU exhibited the highest score for BLEU and CIDEr metrics and DCM-Best1(M) employing

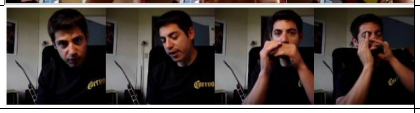
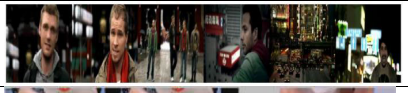
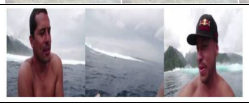
Model	Video frames	Ground Truth & Generated Captions	Explanation
AVSSN		**GT:** a kid pushes a stroller. **G-1 (wo. s-LSTM):** a girl is dancing. **G-2 (wo. AAG):** a bayby is playing. **G-3 (wo. argmax):** a small child is walking. **G-2 (AVSSN):** a girl is pushing a stroller.	A representative sample from MSVD dataset. Improvement in the model performance is observed by analyzing the contribution of the global semantic layer (s-LSTM), Adaptive Attention Gate (AAG), and argmax. The proposed model predicted the verb "pushing" and the noun "stroller" compared to other state of the arts.
SBAT		**GT:** someone is slicing a tomato. **G-1 (Vanilla Transformer):** a woman is slicing an apple. **G-2 (SBAT):** a woman is slicing a tomato.	With the help of redundancy reduction and a better usage of global and local information, SBAT showed better or competitive performance compared with the baseline vanilla transformer for generating caption for a video from MSVD dataset.
SAAT		**GT:** {'a man explains how to solve a rubik s cube', 'a man points at a rubex cube', 'a person discussing how to solve square puzzle', 'a person is solving a rubik s cube', 'a person showing how to solve a rubix cube'} **G-1 (Baseline):** a person is folding a piece of paper. **G-2 (SAAT):** a person is solving a rubik s cube.	Both baseline and SAAT successfully predicted the subject 'person' but the baseline failed to capture the action in the video. Caption generated by SAAT demonstrates the efficacy of the proposed model.
MART		**GT:** A young man wearing a ..... video comes to an end. **G-1 (Vanilla Transformer):** He is sitting down in a chair. He continues ... the camera. **G-2 (T-XL):** A man is seen speaking .... stops playing. **G-2 (MART):** A man is sitting down talking to the ... camera.	Compared to vanilla transformer and transformer-XL, the MART model generated more coherent, less repeated paragraphs while maintaining relevance.
COOT		**GT:** Boil some small pieces of potatoes in water. Mash the potato. Add some butter and salt and stir. Gradually add milk while stirring the potatoes. **G-1 (MART):** Heat up a pan and cook until golden brown. Add onions to the pan. Add flour salt and pepper to the pan. Add rice to the pan and stir. **G-2 (COOT):** Boil the potatoes in water. Add chopped potatoes to the pan. Add butter and mash. Add some milk and mash.	By employing an attention-aware feature aggregation module, 100 times fewer features per video, COOT's video representations encapsulate richer information about the video while being more compact.
MMT	 Alexis: Once we realized Mandy hadn't really cheated on him.	**GT:** • Alexis runs her hand through her hair when Castle is looking at her. (V) • Alexis fixes her hair as she speaks to Castle beside her. (V) • Alexis rubs her hair when Castle is looking at her. (V) • The girl adjust her hair while Castle stares at her. (V) **G-1 (MMT):** • Alexis and castle walk into the room together. (S) • Beckett and castle are talking to each other. (V) • Alexis and castle stand in front of each other as they stand in front of each other. (V+S)	Text inside the dashed box is the subtitle paired with the video clips in TVC dataset, Each GT and generated caption is followed by a tag type, i.e. S, V, and S+V representing subtitle, video or both subtitle and video.
VNS-GRU		**GT:** group of people of singing and walking on the streets. **G-1:** a man is singing and dancing. **G-2:** a group of people are singing and dancing.	G-1 generated while model trained without professional learning and G-2 with Professional learning described video more accurately with advanced words and phrases.
TDConvED		**GT:** a person is slicing an onion. **G-1:** a man is slicing a potato. **G-2:** a man is cutting a vegetable. **G-3:** a man is slicing an onion.	G-1: caption generated using S2VT approach. G-2: caption generated using LSTM-E approach. G-3: Proposed Model generated caption.
OAM		**GT:** a man is lifting the car. **G-1 (O):** a man is lifting a car. **G-2 (OA):** a man is lifting the back of a truck. **G-3 (OAM):** a man is lifting a truck.	O represents the case when only object information is used for caption generation. OA refers to Object and Attention information usage whereas OAM defines full model with Object, Attention and Metric learning Components.
SDN		**GT:** some men having fun and talking about sea. **G-1 (β =0):** a man is talking about a boat. **G-2 ((β =0.7):** a man is talking about the water. **G-3 ((β =1):** a man is talking about the the the the the the the the.	β, a hyperparameter for keeping balance between accuracy and conciseness of the generated caption. **β=0** generates relatively short annotations resulting in incomplete semantics and sentence structure. **β=1** generates redundant words. **β=0.7** generates relatively long sentences without redundancy.
GRU-EVE		**GT:** two teams are playing cricket. **G-1(GRU-MP-CI):** a man is running and falls a ball. **G-2(GRU-EVE$_{hft}$-CI):** a man is playing cricket. **G-3(GRU-EVE):** a man is throwing a cricket ball.	MP-CI: Mean pooling strategy is used to resolve the video temporal dimension. CI represents the joint use of C3D and IRV2, hft stands for Hierarchical Fourier transform and the final proposed model GRU-EVE generated G-3 caption.
EtENet-IRv2		**GT:** a group of elephants are walking. **G-1 (step-1):** a group of people are walking. **G-2 (step-2):** an elephant is walking.	Proposed model has a two stage training setting, firstly, architecture is initialized with the pre-trained encoders and decoders and afterwards the entire network is trained in an end-to-end fashion. G-1 represents the step-1 of proposed model whereas G-2 is generated as step-2.

FIGURE 8. Qualitative Analysis and brief explanation of Models with reference annotations/Ground-Truth(GT) and generated captions(G-1, G-2, G-3).

conditional generative adversarial network, reported excellent scores for METEOR and ROUGE metrics. Table-8 demonstrates results reported on dense captioning datasets like ActivityNet Captions, YouCook-II, TVC, VATEX,and Charades. For ActivityNet captions and YouCook-II datasets, COOT, a cooperative hierarchical transformer model, outperformed all the models for all four metrics BLEU, METEOR, ROUGE, and CIDEr. Similarly, recently proposed HERO on the TVC dataset reported comparable results after a close competition with MMT on the same TVC dataset. Recently proposed transformer-based models exhibited remarkable performance and proved that free from recurrence and solely dependent on self-attention, capable of handling long-term dependency, transformer mechanism handles sequential data

TABLE 7. Video Description - Quantitative Performance Evaluation on MSR-VTT Dataset.

Ref	Approach/ Model	Year	B	M	R	C
[101]	AVSSN	2021	45.5	31.4	64.3	50.6
[102]	SBAT	2020	42.9	28.9	61.5	51.6
[103]	SAAT	2020	40.5	28.2	60.9	49.1
[25]	VNS-GRU	2020	**46**	29.5	63.3	**52.0**
[104]	JSRL-VCT	2019	42.3	29.7	62.8	49.1
[105]	GFN-POS	2019	42.0	28.2	61.6	48.7
[106]	DCM-Best1 (M)	2019	43.8	**34.2**	**65.8**	47.6
[26]	TDConvED (R)	2019	39.5	27.5	-	42.8
[27]	EtENet-IRv2	2019	40.5	27.7	60.6	47.6
[108]	SDN	2019	43.8	28.9	62.4	51.4
[24]	GRU-EVE	2019	38.3	28.4	60.7	48.1
[122]	MM-features	2019	39.2	27.8	59.8	45.7
[109]	OA-BTG	2019	41.4	28.2	-	46.9
[111]	SibNet	2018	40.9	27.5	60.2	47.5
[113]	SeFLA	2018	41.8	-	-	-
[114]	RecNet	2018	39.1	26.6	59.3	42.7
[123]	Lexical- FCN	2017	41.4	28.3	61.1	48.9
[115]	TDDF	2017	37.3	27.8	59.2	43.8

(B:BLEU, M:METEOR, R:ROUGE, C:CIDEr)

TABLE 8. Video Description - Quantitative Performance Evaluation on Miscellaneous Datasets (ANC:ActivityNet Captions, YC2:YouCook-II, VATEX, TVC, Charades, MPII-MD, MVAD, VATEX(En: English, Ch: Chinese)).

Ref	Model (Dataset)	Year	B	M	R	C
[124]	VC-FF (*ANC*)	2021	2.76	7.02	18.16	26.55
[23]	COOT (*ANC*)	2020	**17.43**	**15.99**	**31.45**	**28.19**
[31]	MDVC (*ANC*)	2020	5.83	11.72	-	-
[125]	MART (*ANC*)	2020	9.78	15.57	-	22.16
[126]	I3D+UT (*ANC*)	2019	49	-	-	-
[104]	JSRL-VCT (*ANC*)	2019	1.9	11.3	22.4	44.2
[33]	E2E-MskTr (*ANC*)	2018	2.23	9.56	-	-
[127]	DVC (*ANC*)	2018	1.62	10.33	-	25.24
[23]	COOT(*YC2*)	2020	**17.97**	**19.85**	**37.94**	**57.24**
[125]	MART (*YC2*)	2020	8.0	15.9	-	35.74
[93]	VideoBERT (*YC2*)	2019	4.33	11.94	28.8	0.55
[33]	E2E-MskTr (*YC2*)	2018	1.13	5.9	-	-
[22]	HERO (*TVC*)	2020	**12.35**	**17.64**	**34.16**	**49.98**
[62]	MMT (*TVC*)	2020	10.87	16.91	32.81	45.38
[128]	X-Lin+Tr (*VATEX-En*)	2020	40.7	25.8	53.7	81.4
	X-Lin+Tr (*VATEX-Ch*)		32.6	32.1	56.5	59.5
[112]	Tubes (*Charades*)	2018	31.5	19.1	-	18
[116]	LSTM-TSA (*MPII-MD*)	2017	-	8	-	-
[116]	LSTM-TSA (*M-VAD*)	2017	-	7.2	-	-

(B:BLEU, M:METEOR, R:ROUGE, C:CIDEr, Tr: Transformer)

in a parallel manner allowing accelerated training on large datasets.

IX. CONCLUSION

Video description is an emerging research topic employing cross-modal visuolinguistic modeling for concise and appropriate description generation. These models primarily focus on the compositional structure of convolutional and recurrent neural networks. However, with the technological advancements, these models further accommodate attention mechanisms, reinforcement learning, and transformer mechanism for strength and efficiency. Since the nature of the task is sequence to sequence, it is mainly concerned with the recurrence of the sequential data. One of the issues associated with sequential data processing is the way to deal with the long-term dependencies. The transformer was initially introduced in the NLP domain for text modeling. However, keeping its characteristics of parallelization, recurrence-free, solely dependent on self-attention, accelerated training, space efficiency, and proficiently dealing with the long-term dependencies, it was exploited for both vision and language modeling tasks, i.e., video description. The results for which are tremendous, although in its formative phase. The transformers can be further explored for addressing the concerns hindering progress in this field.

This survey paper explored the benchmark datasets both for pre-training and fine-tuning of video description models. We also investigated the currently available evaluation metrics for assessing the generated captions and emphasized the need for specific and standardized datasets and evaluation metrics to boost the performance. Training the systems with much-complexed/ versatile videos, using subtitles and audio (if available), and providing more reference captions can enhance the system's performance. The introduction of visually diverse and complicated as well as textually dense datasets is a promising research direction.

We expect this survey paper better to understand the video description datasets and evaluation metrics and accommodate researchers in their future explorations and accomplishments.

REFERENCES

[1] *Cisco Annual Internet Report—Cisco Annual Internet Report (2018-2023) White Paper—Cisco*. Accessed: May 4, 2021. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] *10 Youtube Statistics That You Need to Know in 2021*. Accessed: May 4, 2021. [Online]. Available: https://www.oberlo.com/blog/youtube-statistics

[3] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang, "Video in sentences out," in *Proc. 28th Conf. Uncertainty Artif. Intell.*, 2012, pp. 102–112.

[4] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2712–2719.

[5] M. Khan and Y. Gotoh, "Describing video contents in natural language," in *Proc. Workshop Innov. Hybrid*, 2012, pp. 27–35. [Online]. Available: http://dl.acm.org/citation.cfm?id=2388636

[6] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 541–547.

[7] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proc. 25th Int. Conf. Comput. Linguistics*, 2014, pp. 1218–1227.

[8] J. Su, "Study of video captioning problem," 2018. [Online]. Available: https://www.semanticscholar.org/paper/Study-of-Video-Captioning-Problem-Su/511f0041124d8d14bbcdc7f0e57f3bfe13a58e99

[9] C.-J. Fu, G.-H. Li, X.-W. Xu, and K.-X. Dai, "Mining video hierarchical structure for efficient management and access," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 2, Aug. 2006, pp. 13–16.

[10] H. Aradhye, G. Toderici, and J. Yagnik, "Video2Text: Learning to annotate video content," in *Proc. Int. Conf. Data Mining (ICDM)*, 2009, pp. 144–151.

[11] J. Li and H. Qiu. *Comparing Attention-based Neural Architectures for Video Captioning*. Accessed: Feb 10, 2020. [Online]. Available: https://github.com/qijiezhao/pseudo

[12] R. Agyeman, M. Rafiq, H. K. Shin, B. Rinner, and G. S. Choi, "Optimizing spatiotemporal feature learning in 3D convolutional neural networks with pooling blocks," *IEEE Access*, vol. 9, pp. 70797–70805, 2021.

[13] Z. Wang, B. Feng, K. Narasimhan, and O. Russakovsky, "Towards unique and informative captioning of images," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, vol. 12352. Glasgow, U.K.: Springer, Aug. 2020, pp. 629–644. [Online]. Available: https://collaborate.princeton.edu/en/publications/towards-unique-and-informative-captioning-of-images

[14] E. Takmaz, S. Pezzelle, L. Beinborn, and R. Fernández, "Generating image descriptions via sequential cross-modal alignment guided by human gaze," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4664–4677.

[15] H. Rampal and A. Mohanty, "Efficient CNN-LSTM based image captioning using neural network compression," 2020, *arXiv:2012.09708*. [Online]. Available: http://arxiv.org/abs/2012.09708

[16] Z. Meng, L. Yu, N. Zhang, T. Berg, B. Damavandi, V. Singh, and A. Bearman, "Connecting what to say with where to look by modeling human attention traces," 2021, *arXiv:2105.05964*. [Online]. Available: http://arxiv.org/abs/2105.05964

[17] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," 2021, *arXiv:2101.06462*. [Online]. Available: http://arxiv.org/abs/2101.06462

[18] E. Bugliarello and D. Elliott, "The role of syntactic planning in compositional image captioning," 2021, *arXiv:2101.11911*. [Online]. Available: http://arxiv.org/abs/2101.11911

[19] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "VisualGPT: Data-efficient adaptation of pretrained language models for image captioning," 2021, *arXiv:2102.10407*. [Online]. Available: http://arxiv.org/abs/2102.10407

[20] L. Chen, Z. Jiang, J. Xiao, and W. Liu, "Human-like controllable image captioning with verb-specific semantic roles," 2021, *arXiv:2103.12204*. [Online]. Available: http://arxiv.org/abs/2103.12204

[21] G. Xu, S. Niu, M. Tan, Y. Luo, Q. Du, and Q. Wu, "Towards accurate text-based image captioning with content diversity exploration," 2021, *arXiv:2105.03236*. [Online]. Available: http://arxiv.org/abs/2105.03236

[22] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "HERO: Hierarchical encoder for video+language omni-representation pre-training," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2046–2065.

[23] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "COOT: Cooperative hierarchical transformer for video-text representation learning," 2020, *arXiv:2011.00597*. [Online]. Available: http://arxiv.org/abs/2011.00597

[24] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12479–12488.

[25] H. Chen, J. Li, and X. Hu, "Delving deeper into the decoder for video captioning," 2020, *arXiv:2001.05614*. [Online]. Available: http://arxiv.org/abs/2001.05614

[26] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, "Temporal deformable convolutional encoder-decoder networks for video captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8167–8174.

[27] S. Olivastri, G. Singh, and F. Cuzzolin, "End-to-end video captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9. [Online]. Available: https://zhuanzhi.ai/paper/004e3568315600ed58e6a699bef3cbba

[28] L. Li and B. Gong, "End-to-end video captioning with multitask reinforcement learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 339–348.

[29] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8393–8400.

[30] Z. Fang, T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "Video2Commonsense: Generating commonsense descriptions to enrich video captioning," 2020, *arXiv:2003.05162*. [Online]. Available: http://arxiv.org/abs/2003.05162

[31] V. Iashin, "Multi-modal dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 958–959.

[32] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[33] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8739–8748.

[34] M. Rafiq, G. Rafiq, R. Agyeman, S.-I. Jin, and G. Choi, "Scene classification for sports video summarization using transfer learning," *Sensors*, vol. 20, no. 6, p. 1702, 2020.

[35] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 4565–4574.

[36] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 706–715.

[37] J. Varghese and K. N. Nair, "An algorithmic approach for general video summarization," in *Proc. 5th Int. Conf. Adv. Comput. Commun.*, 2015, pp. 7–11.

[38] P. Gunawardena, H. Sudarshana, O. Amila, R. Nawaratne, D. Alahakoon, A. S. Perera, and C. Chitraranjan, "Interest-oriented video summarization with keyframe extraction," in *Proc. 19th Int. Conf. Adv. Emerg. Regions*, vol. 1, 2019, pp. 1–8.

[39] M. Fei, W. Jiang, and W. Mao, "Creating personalized video summaries via semantic event detection," *J. Ambient Intell. Hum. Comput.*, vol. 4, pp. 1–12, Apr. 2018, doi: 10.1007/s12652-018-0797-0.

[40] A. Yoshitaka and K. Sawada, "Personalized video summarization based on behavior of viewer," in *Proc. 8th Int. Conf. Signal Image Technol. Internet Based Syst.*, Nov. 2012, pp. 661–667.

[41] H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation methods," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–13, Jan. 2020.

[42] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 1–37, 2019.

[43] J. Park, C. Song, and J. H. Han, "A study of evaluation metrics and datasets for video captioning," in *Proc. 2nd Int. Conf. Intell. Inform. Biomed. Sci.*, Jan. 2018, pp. 172–175.

[44] M. Amaresh and S. Chitrakala, "Video captioning using deep learning: An overview of methods, datasets and metrics," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2019, pp. 656–661.

[45] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," *Frontiers Multimedia Res.*, vol. 4, pp. 3–29, Dec. 2017.

[46] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2011, pp. 190–200.

[47] M. Rohrbach and M. Planck, "A database for fine grained activity detection of cooking activities," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1194–1201.

[48] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Oct. 2013, pp. 2634–2641.

[49] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.

[50] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, "Coherent multi-sentence video description with variable level of detail," in *Proc. 36th GCPR*, vol. 8753. Münster, Germany: Springer, 2014, pp. 184–195. [Online]. Available: https://dblp.uni-trier.de/db/conf/dagm/gcpr2014.html#RohrbachRQFPS14

[51] L. Zhou and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2016, pp. 7590–7598.

[52] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3202–3212.

[53] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," 2015, *arXiv:1503.01070*. [Online]. Available: http://arxiv.org/abs/1503.01070

[54] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, "Grounded video description," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6571–6580.
[55] S. Gella, M. Lewis, and M. Rohrbach, "A dataset for telling the stories of social media videos," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 968–974.
[56] G. A. Sigurdsson, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," 2016, *arXiv:1604.01753*. [Online]. Available: https://arxiv.org/abs/1604.01753
[57] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 5288–5296, 2016.
[58] K. H. Zeng, T. H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *Proc. 14th Eur. Conf.*, vol. 9906. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 609–625. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46475-6_38
[59] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4581–4591.
[60] S. Zhang, Z. Tan, J. Yu, Z. Zhao, K. Kuang, J. Liu, J. Zhou, H. Yang, and F. Wu, "Poet: Product-oriented video captioner for E-commerce," 2020, *arXiv:2008.06880*. [Online]. Available: http://arxiv.org/abs/2008.06880
[61] G. Huang, B. Pang, Z. Zhu, C. Rivera, and R. Soricut, "Multimodal pre-training for dense video captioning," 2020, *arXiv:2011.11760*. [Online]. Available: http://arxiv.org/abs/2011.11760
[62] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVR: A large-scale dataset for video-subtitle moment retrieval," in *Proc. 16th Eur. Conf.*, vol. 12366. Glasgow, U.K.: Springer, Aug. 2020, pp. 447–463. [Online]. Available: https://www.springerprofessional.de/tvr-a-large-scale-dataset-for-video-subtitle-moment-retrieval/18577886
[63] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Proc. 12th Eur. Conf. Comput. Vis.*, vol. 7572. Florence, Italy: Springer, Oct. 2012, pp. 144–157. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-33718-5_11
[64] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 12, Oct. 2015, pp. 961–970.
[65] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950
[66] S. Abu-el haija, J. Lee, P. Natsev, and G. Toderici, "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*. [Online]. Available: https://arxiv.org/abs/1609.08675
[67] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 228–231. [Online]. Available: http://acl.ldc.upenn.edu/W/W05/W05-09.pdf#page=75
[68] V. Tech, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," 2014, *arXiv:1411.5726*. [Online]. Available: https://arxiv.org/abs/1411.5726
[69] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. 14th Eur. Conf.*, vol. 9909. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 382–398. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46454-1_24
[70] N. Sharif, L. White, M. Bennamoun, and S. A. Ali Shah, "Learning-based composite metrics for improved caption evaluation," in *Proc. Student Res. Workshop*, 2018, pp. 14–20.
[71] Y. Zhang and S. Vogel, "Significance tests of automatic machine translation evaluation metrics," *Mach. Transl.*, vol. 24, no. 1, pp. 51–65, Mar. 2010.
[72] G. Wentzel, "Funkenlinien im Röntgenspektrum," *Ann. Phys.*, vol. 371, no. 23, pp. 437–461, 1922.
[73] C.-Y. Lin and E. Hovy, "Manual and automatic evaluation of summaries," in *Proc. Workshop Autom. Summarization*, 2002, pp. 25–26. [Online]. Available: https://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85
[74] H. Saggion, T. Poibeau, H. Saggion, T. Poibeau, A. Text, S. Past, and T. Future, "Automatic text summarization: Past, present and future," in *Proc. Multi-Source, Multilingual Inf. Extraction Summarization*, 2016, pp. 3–21.
[75] M. J. Kusner, Y. Sun, N. I. Kolkin, Q. K. Weinberger, and S. K. Haldar, "From word embeddings to document distances matt," *Ironmaking Steelmaking*, vol. 44, no. 7, pp. 526–531, 2017.
[76] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 199–209.
[77] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, and M. Nagata, "SODA: Story oriented dense video captioning evaluation framework," in *Proc. 16th ECCV*, vol. 6. Glasgow, U.K.: Springer, 2020, pp. 517–531. [Online]. Available: https://dblp.uni-trier.de/db/conf/eccv/eccv2020-6.html#FujitaHKON20
[78] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *J. Document.*, vol. 60, no. 5, pp. 503–520, 2004.
[79] G. Doddington, "Automatic evaluation of machine translation quality using N-gram co-occurrence statistics," in *Proc. 2nd Int. Conf. Hum. Lang. Technol. Res.*, 2002, pp. 138–145.
[80] A. Lavie, K. Sagae, and S. Jayaraman, "The significance of recall in automatic metrics for MT evaluation," in *Proc. 6th Conf. Assoc. Mach. Transl. Amer. (AMTA)*, vol. 3265. Washington, DC, USA: Springer, Sep./Oct. 2004, pp. 134–143. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-30194-3_16
[81] C. Fellbaum, "WordNet," in *The Encyclopedia of Applied Linguistics*. Hoboken, NJ, USA: Wiley, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal1285
[82] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.
[83] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 452–457.
[84] C.-Y. Lin, E. Hovy, and M. Rey, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proc. Hum. Lang. Technol. Conf. North Amer. Assoc.*, 2003, pp. 150–157.
[85] T. Demeester, T. Rocktäschel, and S. Riedel, "Lifted rule injection for relation embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1389–1399.
[86] C. L. Zitnick and D. Parikh, "Bringing semantics into focus using visual abstraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3009–3016.
[87] J. Giménez and L. Márquez, "Linguistic features for automatic evaluation of heterogenous MT systems," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 256–264.
[88] L. D. Ellebracht, A. Ramisa, P. S. Madhyastha, J. Cordero-Rama, F. Moreno-Noguer, and A. Quattoni, "Semantic tuples for evaluation of image to sentence generation," in *Proc. 4th Workshop Vis. Lang.*, 2015, pp. 18–28.
[89] H. Xu, G. Ghosh, P.-Y. Huang, P. Arora, M. Aminzadeh, C. Feichtenhofer, F. Metze, and L. Zettlemoyer, "VLM: Task-agnostic video-language model pre-training for video understanding," 2021, *arXiv:2105.09996*. [Online]. Available: http://arxiv.org/abs/2105.09996
[90] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "UniVL: A unified video and language pre-training model for multimodal understanding and generation," 2020, *arXiv:2002.06353*. [Online]. Available: http://arxiv.org/abs/2002.06353
[91] L. Zhu and Y. Yang, "ActBERT: Learning global-local video-text representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8743–8752.
[92] Y. Pan, Y. Li, J. Luo, J. Xu, T. Yao, and T. Mei, "Auto-captions on GIF: A large-scale video-sentence dataset for vision-language pre-training," 2020, *arXiv:2007.02375*. [Online]. Available: http://arxiv.org/abs/2007.02375
[93] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7463–7472.
[94] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," 2019, *arXiv:1906.05743*. [Online]. Available: http://arxiv.org/abs/1906.05743
[95] M.-W. C. Kenton, L. Kristina, and J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," 1953, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805

[96] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*. [Online]. Available: http://arxiv.org/abs/1908.02265

[97] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[98] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, pp. 10384–10394.

[99] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2630–2640.

[100] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba, "Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 187–203, Jan. 2021.

[101] J. Perez-Martin, B. Bustos, and J. Perez, "Attentive visual semantic specialized network for video captioning," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5767–5774.

[102] T. Jin, S. Huang, M. Chen, Y. Li, and Z. Zhang, "SBAT: Video captioning with sparse boundary-aware transformer," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 630–636.

[103] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13093–13102.

[104] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8917–8926.

[105] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable video captioning with pos sequence guidance based on gated fusion network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2641–2650.

[106] H. Xiao and J. Shi, "Diverse video captioning through latent variable expansion," 2019, *arXiv:1910.12019*. [Online]. Available: http://arxiv.org/abs/1910.12019

[107] R. J. Babariya and T. Tamaki, "Meaning guided video captioning," 2019, *arXiv:1912.05730*. [Online]. Available: https://arxiv.org/abs/1912.05730

[108] H. Chen, K. Lin, A. Maye, J. Li, and X. Hu, "A semantics-assisted video captioning model trained with scheduled sampling," 2019, *arXiv:1909.00121*. [Online]. Available: http://arxiv.org/abs/1909.00121

[109] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," 2019, *arXiv:1906.04375*. [Online]. Available: http://arxiv.org/abs/1906.04375

[110] J. Lee, Y. Lee, S. Seong, K. Kim, S. Kim, and J. Kim, "Capturing long-range dependencies in video captioning," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1880–1884.

[111] S. Liu, Z. Ren, and J. Yuan, "SibNet: Sibling convolutional encoder for video captioning," in *Proc. ACM Multimedia Conf.*, 2018, pp. 1425–1434.

[112] B. Zhao, X. Li, and X. Lu, "Video captioning with tube features," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1177–1183.

[113] S. Lee and I. Kim, "Multimodal feature learning for video captioning," *Math. Problems Eng.*, vol. 2018, Feb. 2018, Art. no. 3125879.

[114] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7622–7631.

[115] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Task-driven dynamic fusion: Reducing ambiguity in video description," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2017, pp. 6250–6258.

[116] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 984–992.

[117] D. Wang and D. Song, "Video captioning with semantic information from the knowledge base," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Aug. 2017, pp. 224–229.

[118] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4594–4602.

[119] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4534–4542.

[120] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 1–11.

[121] M. Koupaee and W. Yang Wang, "WikiHow: A large scale text summarization dataset," 2018, *arXiv:1810.09305*. [Online]. Available: http://arxiv.org/abs/1810.09305

[122] M. Hammad, M. Hammad, and M. Elshenawy, "Characterizing the impact of using features extracted from pre-trained models on the quality of video captioning sequence-to-sequence models," 2019, *arXiv:1911.09989*. [Online]. Available: http://arxiv.org/abs/1911.09989

[123] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5159–5167.

[124] M. Hosseinzadeh and Y. Wang, "Video captioning of future frames," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 980–989.

[125] J. Lei, L. Wang, Y. Shen, D. Yu, T. Berg, and M. Bansal, "MART: Memory-augmented recurrent transformer for coherent video paragraph captioning," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2603–2614.

[126] M. Bilkhu, S. Wang, and T. Dobhal, "Attention is all you need for videos: Self-attention based video summarization using universal transformers," 2019, *arXiv:1906.02792*. [Online]. Available: http://arxiv.org/abs/1906.02792

[127] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly localizing and describing events for dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7492–7500.

[128] X. Zhu, L. Guo, P. Yao, S. Lu, W. Liu, and J. Liu, "Vatex video captioning challenge 2020: Multi-view features and hybrid reward strategies for video captioning," 2019, *arXiv:1910.11102*. [Online]. Available: http://arxiv.org/abs/1910.11102

MUHAMMAD RAFIQ received the M.S. degree in electronics engineering from the International Islamic University, Pakistan, in 2008. He is currently pursuing the Ph.D. degree with Yeungnam University, South Korea. He has extensive industry experience with a background in database, business applications, and industrial technical solutions.

His research interests include smart sensors, computer vision, medical image analysis, and video description using deep learning.

GHAZALA RAFIQ received the B.Sc. degree in mathematics from Punjab University, Lahore, Pakistan, in 2000, and the master's degree in computer science, in 2002. She is currently pursuing the Ph.D. degree with the Data Sciences Laboratory, Department of Information and Communication Engineering, Yeungnam University, Republic of Korea. She has over 15 years of industry experience.

Her research interests include deep-learning applications, video description, reinforcement learning, natural language processing, computer vision, and pattern recognition.

GYU SANG CHOI (Member, IEEE) received the Ph.D. degree in computer science and engineering from Pennsylvania State University. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been with Yeungnam University, where he is currently an Assistant Professor. His research interests include embedded systems, storage systems, parallel and distributed computing, supercomputing, cluster-based web servers, and data centers.

He is currently working on embedded systems and storage systems, while his prior research has been mainly focused on improving the performance of clusters. He is a member of the ACM.