IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 December 2023, accepted 11 January 2024, date of publication 24 January 2024, date of current version 11 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357980

APPLIED RESEARCH

Combinatorial Analysis of Deep Learning and Machine Learning Video Captioning Studies: A Systematic Literature Review

TANZILA KEHKASHAN[1], **ABDULLAH ALSAEEDI**[2], **WAEL M. S. YAFOOZ**[2], **NOR AZMAN ISMAIL**[1], **AND ARAFAT AL-DHAQM**[3]

[1]Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[2]Computer Science Department, College of Computer Science and Engineering, Taibah University, Madina 42353, Saudi Arabia
[3]Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Perak 32610, Malaysia

Corresponding authors: Tanzila Kehkashan (tanzila@graduate.utm.my) and Wael M. S. Yafooz (waelmohammed@hotmail.com)

ABSTRACT Recent improvements formulated in the area of video captioning have brought rapid revolutions in its methods and the performance of its models. Machine learning and deep learning techniques are both employed in this regard. However, there is a lack of tracing the latest studies and their remarkable results. Although several studies have been proposed employing the ML and DL algorithms in different other areas, there is no systematic review utilizing the video captioning task. This study aims to examine, evaluate, and synthesize the primary studies into a thorough Systematic Literature Review (SLR) that provides a general overview of the methods used for video captioning. We performed the SLR to determine the research problems under which machine learning models were preferred over the deep learning models and vice versa. We collected a total of 1,656 studies retrieved from four electronic databases; Scopus, WoS, IEEE Xplore, and ACM, based on our search string from which 162 published studies passed the selection criteria related to one primary and two secondary research questions after a systematic process. Moreover, insufficient data collection and inefficient comparison of results are common issues identified during the review process. We conclude that the 2D/3D CNN for video feature extraction and LSTM for caption generation, METEOR and BLEU performance evaluation tools, and MSVD dataset are most frequently employed for video captioning. Our study is the pioneer in comparing the implementation of ML and DL algorithms employing the video captioning area. Thus, our study will accelerate the critical assessment of the state-of-the-art in other research fields of video analysis and human-computer interaction.

INDEX TERMS Deep learning, machine learning, performance evaluation metrics, video analysis.

I. INTRODUCTION

Video captioning, an area of Computer Vision (CV) and Natural Language Processing (NLP) has gained significant research focus in recent years. The main goal of computer vision is to convey the visual information in a video in natural language, which not only gives a comprehensive overview of the video but also organizes the visual data into sentences with good grammar and logical structures with decent words. The idea behind is somehow more detailed, more complex, and more challenging than video subtitling as the subtitling involves only the conversion of audio spoken in the video into natural language text. Video captioning, on the other hand, aims to understand every entity in the video sequences and describe it in grammatically accurate natural language. These entities include scenes, objects, actions, and all the interactions among these entities because this description is not understandable for humans without interaction details.

The scope of video captioning has significantly expanded in the modern era as a result of the tremendous advancements

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

in technology and captioning algorithms. This systematic study ponders light on different aspects of video captioning studies proposed so far. The development of video captioning creates a wide range of new options. Advancements in video captioning can be witnessed in numerous application domains like sign-language recognition, video surveillance, human-robot interaction, visually or hearing-impaired assistance, video indexing, and many more [1]. We anticipate having the same kind of interactions with robots that we have with people soon. There are two distinct steps in the process of captioning videos; visual understanding and caption generation. Four basic modalities— visual, audio, motion, and semantic—are necessary to build a video interpretation paradigm. Using cutting-edge techniques, numerous researchers have had success in extracting the features from diverse modalities.

A combinatorial analysis of deep learning and machine learning video captioning studies can provide valuable insights into the state of the art in this field. By comparing and contrasting the different approaches used in these studies, it is possible to identify the strengths and weaknesses of each approach and to determine the best approaches for video captioning. One aspect of a combinatorial analysis is to classify the studies based on the type of deep learning or machine learning models used. For example, studies may use convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformers. This can provide insights into the most commonly used models, and the performance of each model. Another aspect is to classify the studies based on the type of datasets used for training and evaluating the models. For example, studies may use video captions, subtitles, or audio-visual data. This can provide insights into the most suitable data for different models, and the impact of the data on the performance of the models. A further aspect is to classify the studies based on the evaluation metrics used. For example, studies may use metrics such as BLEU, ROUGE, or METEOR. This can provide insights into the most commonly used metrics, and the suitability of each metric for different models and data.

The issue of efficiently describing material using images has been the subject of numerous research previously [2], [3]. The challenge of characterizing actions through videos, however, is more challenging to increase accuracy in video feature extraction [4]. Numerous approaches have been employed for the captioning of videos over the past few years. Early studies of video captioning approaches presented the visual themes in the video with hand-crafted features and template-based caption generation. These template-based or rule-based SVO approaches use the Subject (S), Verb (V), and Object (O) triplets and generate the caption based on predefined sentence templates [5], [6], [7]. These classical methods of SVO were highly dependent on the templates and used some fixed syntactic structure of sentences. However, this approach is incapable of describing all the entities of open-domain videos using hand-crafted features because it is not practical due to its computational complexity. In recent years, the statistical or machine learning approach was employed using Statistical Machine Translation (SMT) techniques [8] where intermediate semantic label representation was used to generate associated annotations to cope with the issues of large-scale and open-domain video datasets. There is a long list of video captioning studies based on statistical approaches from traditional classifiers [9] to weakly supervised [10] and fully supervised [11]. However, to adequately capture the relationship between visual features and textual descriptions, these strategies worked insufficiently. Recent research developments in the CV and NLP areas using deep learning approaches have leveraged the concept of generating sentences from video pixels. These studies employed the DL-based architectures [12] that encode the visual features using different variations of CNN and decode these features into generated captions using different language models i. e., RNN, LSTM, GRU.

Recent publications have unavoidably showcased the most cutting-edge approaches to various video captioning techniques. A recent review of video captioning was conducted by Jain et. al. [4] that emphasizes various features of different video captioning techniques and datasets. The article offers a summary of the many deep-learning architectures used for video captioning. According to the authors, the models that were primarily used, with a few small adjustments and various attention and encoder-decoder framework combinations led to either a rise or a decline in the results. Islam et. al. [13] surveyed various classical and deep learning methods already in use. The review covers benchmark datasets, prominent assessment criteria, and architectural frameworks. The authors also address the applications and limitations of video captioning methods. The information already available about video description is also examined in a literature review by Aafaq et al. [1]. In addition to comparing the domains of benchmark datasets, number of classes, and repository sizes, this study also examines the benefits and drawbacks of different evaluation measures, including SPICE, CIDEr, ROUGE, BLEU, METEOR, and WMD. Moreover, it highlights the cutting-edge approaches of the video description with an emphasis on deep learning models. The paper also summarizes the benchmark outcomes of different methodologies on each dataset of video descriptions. It was discovered that the heterogeneity in unrestricted open-domain videos cannot be handled by the classical approaches as a result of the release of massive datasets. Almost every study agrees that there is no specific performance evaluation metric available. Although metrics for image captioning and machine translation have been borrowed in the evaluation of video captions, there is no specific metric designed to gauge the effectiveness of video descriptions.

This systematic study will help the researchers to decide the methods for captioning the videos for their specific domain. This document provides a thorough review of the most recent studies on video captioning published so far. Various ML and DL algorithms have been employed for video captioning tasks so far. However, no synthesis of these studies has been reported to provide a systematic review of video captioning. Thus, a systematic study is lacking yet in this domain. Our goal is to close this disparity by conducting an SLR to focus on the ML and DL algorithms applied in video captioning, evaluation parameters, and datasets used along with the challenges encountered by answering our research questions. So, the key contributions of this study are:

i. The conducted study is the first systematic study in the area of video captioning.
ii. We critically analyzed 162 relevant primary studies to address our research question.
iii. We identified the research gaps and explored future research directions in the area of video captioning.

The organization of this study's structure is as follows: The employed research methodology is thoroughly described in section II. Section III discusses the results and interpretation comprehensively. The study's final remarks are included in section IV. Lastly, section 5 describes some important highlights regarding the future aspects. We have followed the PRISMA 2020 checklist [14] to address all the necessary sections of this systematic literature review. Further information on the studies used in this SLR is presented in Appendix A at the end of this SLR.

A. METHODOLOGY

A systematic literature review is a technique used to recognize, and critically assess all the available research studies related to some specific research question or topic of interest [15].

This systematic review's goal is to offer a discussion to summarize the existing research area of video captioning, find out its research gaps, and explore potential future research. It highlights all the existing primary research studies using some explicit methods to answer our primary research question and thus forms a secondary research study. The phases involved in this systematic review are planning, identification, selection, assessment, and reporting (see Figure 1) respectively discussed below in detail:

B. REVIEW PLAN

A systematic study is enviable because it concludes and summarizes the literature covering some scope that is not possible from individual studies [15]. Before conducting the study, it was verified whether there was no prior study published covering the same research scope as ours. After the confirmation, the following steps were carried out in planning the review phase:

1) REVIEW PROTOCOL

The review protocol presents the outline of all the components of the review along with some more essential information. An inadequate protocol may lead to biased study selection results, that's why we have carefully developed the review protocol. Due to the iterative development and evaluation, our review protocol shows the planned methods of our review. It was prepared as a guide before the actual review would be conducted. PRISMA-P Checklist [16] was followed to develop the protocol and is described in the Appendix A highlighting the important points of the study, i.e., RQs, inclusion/exclusion criteria, data extraction, and synthesis.

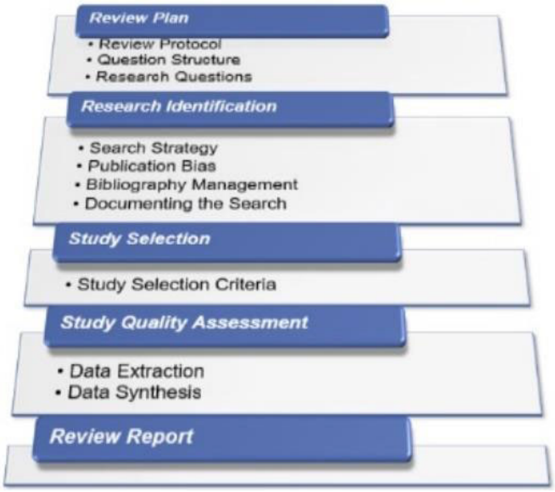

FIGURE 1. Systematic review framework.

TABLE 1. PICO framework for research questions.

Problem	video captioning
Intervention	deep learning method
Comparison	machine learning method
Outcomes	effectiveness

Question Structure: Formulating the structure of the research question is the fundamental process for research to be conducted. We have used the PICO (Problem, Intervention, Comparison, Outcome) framework [17] to formulate our research questions for this systematic review.

We adopted the PICO framework to join the several elements of our research question. It is a general rule of PICO to reflect the research question in the study title or the abstract of the study and we have done it in both.

2) RESEARCH QUESTIONS

Various methods, datasets, and performance evaluation metrics have been employed to identify the elements involved in single-sentence and dense video captioning. We conducted this review to answer the following primary research question to examine the studies related to single-sentence and dense video captioning methods:

RQ 1: *What evidence indicates that deep learning methods are more commonly utilized than machine learning methods in video captioning?*

Since all these studies used different datasets and performance metrics, we formulated the following secondary questions to assist this review:

RQ 2: *Which video captioning datasets relative to different parameter characteristics are used preferably for the benchmarking?*

RQ 3: *Which performance evaluation metric is most appropriate for the studies of video captioning models relative to parameter characteristics?*

The research questions were developed to fully cover the objectives of this systematic review – from identifying the methods used, to the characteristics of under-study datasets. We also intended to investigate the issue of performance evaluation metrics used for video captioning. These RQs helped us to develop our research strategy and identify inclusion and exclusion criteria.

C. RESEARCH IDENTIFICATION

Using the following objective search strategy, the goal of this systematic literature review is to look for as many primary studies as possible that are pertinent to our research question.

1) SEARCH STRATEGY

After the formulation of the research questions, the following steps were adopted to design the search strategy for identifying pertinent studies:

i. Using the PICO approach [17], the major search terms were derived from the study questions. Important concepts of the PICO framework were translated into the search terms. Listed below are the key search terms: ·*Problem:* video, captioning. ·*Intervention:* deep, learning, method. ·*Comparison:* machine, learning, method. ·*Outcomes:* influence, of, video, captioning, model.

ii. Identified alternative spellings and synonyms for search terms using the following concept grid. Similar terms were added to the grid under its related concept.

iii. Checked the keywords in the relevant papers.

iv. To broaden our search, we used the Boolean OR and incorporated alternative spellings and synonyms having similar concepts.

v. To narrow down our search, we used the Boolean AND and linked the major terms from step 1 having different concepts.

Below search-strategy was tailored to four established databases; i.e., Scopus, Web of Science, IEEE Xplore, and ACM Digital Library:

video caption OR visual caption* OR video description OR visual description OR describing video OR video-to-text*

AND

deep learning OR deep-learning OR neural network

AND

method OR approach OR technique OR mechanism OR model

2) PUBLICATION BIAS

Publication bias of a systematic review is the tendency of authors to include some specific studies only and it affects the quality of the research. This is a common problem, as it can skew the results and conclusions of the review. To avoid publication bias in our SLR of video captioning studies, we adopted a rigorous and comprehensive approach. Here are the steps that were followed:

i. *Searched from multiple databases:* we searched from multiple databases, including both engineering and computing sources, to ensure a comprehensive and representative sample of the available literature.

ii. *Used comprehensive search strategy:* We developed a comprehensive search strategy that covers a wide range of keywords and phrases related to video captioning and machine learning/deep learning and used advanced search filters to exclude irrelevant studies.

iii. *Included grey literature:* We have included conference proceedings and technical reports, in addition to published journal articles, to ensure that all relevant studies are included.

iv. *Multiple reviewers for studies screening:* We assigned multiple reviewers to independently screen the studies, and used a consensus approach to ensure that all relevant studies are included.

v. *Assessed the quality of the studies:* We assessed the quality of the studies based on our established inclusion and exclusion criteria, such as the study design, and the validity and reliability of the data to synthesize the results of the studies, so that we can provide a comprehensive and robust evaluation of the state of the art in video captioning. These criteria were formulated by the first author and were validated by all the remaining authors.

By following these steps, we minimized the risk of publication bias and ensured a rigorous and comprehensive SLR of video captioning studies.

3) BIBLIOGRAPHY MANAGEMENT PACKAGE USED FOR DOCUMENT RETRIEVAL

We used Endnote software to manage and organize a large number of search results as references and citations systematically and efficiently. Hundreds of irrelevant journal and conference proceedings studies were returned. Numerous duplicate searches among our selected four databases were retrieved. By using Endnote, we removed duplicate records and created a library of references representing our initial database.

4) DOCUMENTING THE SEARCH

The fundamental concept of search string definition is to join the synonyms of keywords with OR connector and concatenate different groups of words with AND. Table 3 contains the documentation of the search for each database selected.

D. STUDY SELECTION

We looked for the searches spanning from the inception of the database until October 07, 2023, and included journal articles, and conference papers using the search strings shown in Table 3. Figure 2 describes that a total of 1,656 records were retrieved using the four literature repositories. The largest

TABLE 2. Concept grid.

Concept 1: video captioning	Concept 2: deep learning	Concept 3: method
video caption visual caption video description visual description describing video video-to-text	deep learning deep-learning neural network	Method Approach Technique Mechanism Model

TABLE 3. Studies search strategies documentation for selected databases.

	Search String	Search date	Coverage date provided by database	Total publication found	Relevant publications
Scopus	TITLE-ABS-KEY (({video caption} OR {visual caption} OR {video description} OR {visual description} OR {describing video} OR {video-to-text}) AND ({deep learning} OR {deep-learning} OR {neural network}) AND (method OR approach OR technique OR mechanism OR model))	Oct 07, 2023	2002-2023	200	76
Web of Science	(((TI=(((video caption* OR visual caption* OR video description OR visual description OR describing video OR video-to-text) AND (deep learning OR deep-learning OR neural network) AND (method OR approach OR technique OR mechanism OR model)))) OR AB=(((video caption* OR visual caption* OR video description OR visual description OR describing video OR video-to-text) AND (deep learning OR deep-learning OR neural network) AND (method OR approach OR technique OR mechanism OR model)))) AND AK=(((video caption* OR visual caption* OR video description OR visual description OR describing video OR video-to-text) AND (deep learning OR deep-learning OR neural network) AND (method OR approach OR technique OR mechanism OR model))))	Oct 07, 2023	2017-2023	25	11
IEEE Xplore	(("Document Title":video caption* OR "Document Title":visual caption* OR "Document Title":video description OR "Document Title":visual description OR "Document Title":describing video OR "Document Title":video-to-text) AND ("Document Title":deep learning OR "Document Title":deep-learning OR "Document Title":neural network) AND ("Document Title":method OR "Document Title":approach OR "Document Title":technique OR "Document Title":mechanism OR "Document Title":model)) OR (("Abstract":video caption* OR "Abstract":visual caption* OR "Abstract":video description OR "Abstract":visual description OR "Abstract":describing video OR "Abstract":video-to-text) AND ("Abstract":deep learning OR "Abstract":deep-learning OR "Abstract":neural network) AND ("Abstract":method OR "Abstract":approach OR "Abstract":technique OR "Abstract":mechanism OR "Abstract":model)) OR (("Index Terms":video caption* OR "Index Terms":visual caption* OR "Index Terms":video description OR "Index Terms":visual description OR "Index Terms":describing video OR "Index Terms":video-to-text) AND ("Index Terms":deep learning OR "Index Terms":deep-learning OR "Index Terms":neural network) AND ("Index Terms":method OR "Index Terms":approach OR "Index Terms":technique OR "Index Terms":mechanism OR "Index Terms":model))	Oct 07, 2023	1989 – 2023	1,310	73
ACM Digital	(Title: “video caption*” OR Title: “visual caption*” OR Title: “video description OR Title: “visual description” OR Title: “describing video” OR “Title: “video-to-text”) AND (Abstract: “video caption*” OR Abstract: “visual caption*” OR Abstract: “video description” OR Abstract: “visual description” OR Abstract: “describing video” OR Abstract: “video-to-text”) AND (Abstract: “deep learning” OR Abstract: “deep-learning” OR Abstract: “neural network”) AND (Abstract: method OR Abstract: approach OR Abstract: technique OR Abstract: mechanism OR Abstract: model) AND (Keywords: “video caption*” OR Keywords: “visual caption*” OR Keywords: “video description” OR Keywords: “visual description” OR Keywords: “describing video” OR Keywords: “video-to-text”) AND (Keywords: “deep learning” OR Keywords: “deep-learning” OR Keywords: “neural network”) AND (Keywords: method OR Keywords: approach OR Keywords: technique OR Keywords: mechanism OR Keywords: model)	Oct 07, 2023	1989-2023	121	5

number of studies were found using IEEE Xplore. It is quite obvious that a study can be indexed in many databases, so we looked for duplicate video captioning studies and removed the 84 duplicate studies from either of the databases. After obtaining the initial 1,572 studies, we employed the following three-stage screening process utilized by Stefana et. al. [18]:

Screening Stage 1 – Title analysis: In this stage, only the titles of retrieved studies were analyzed.

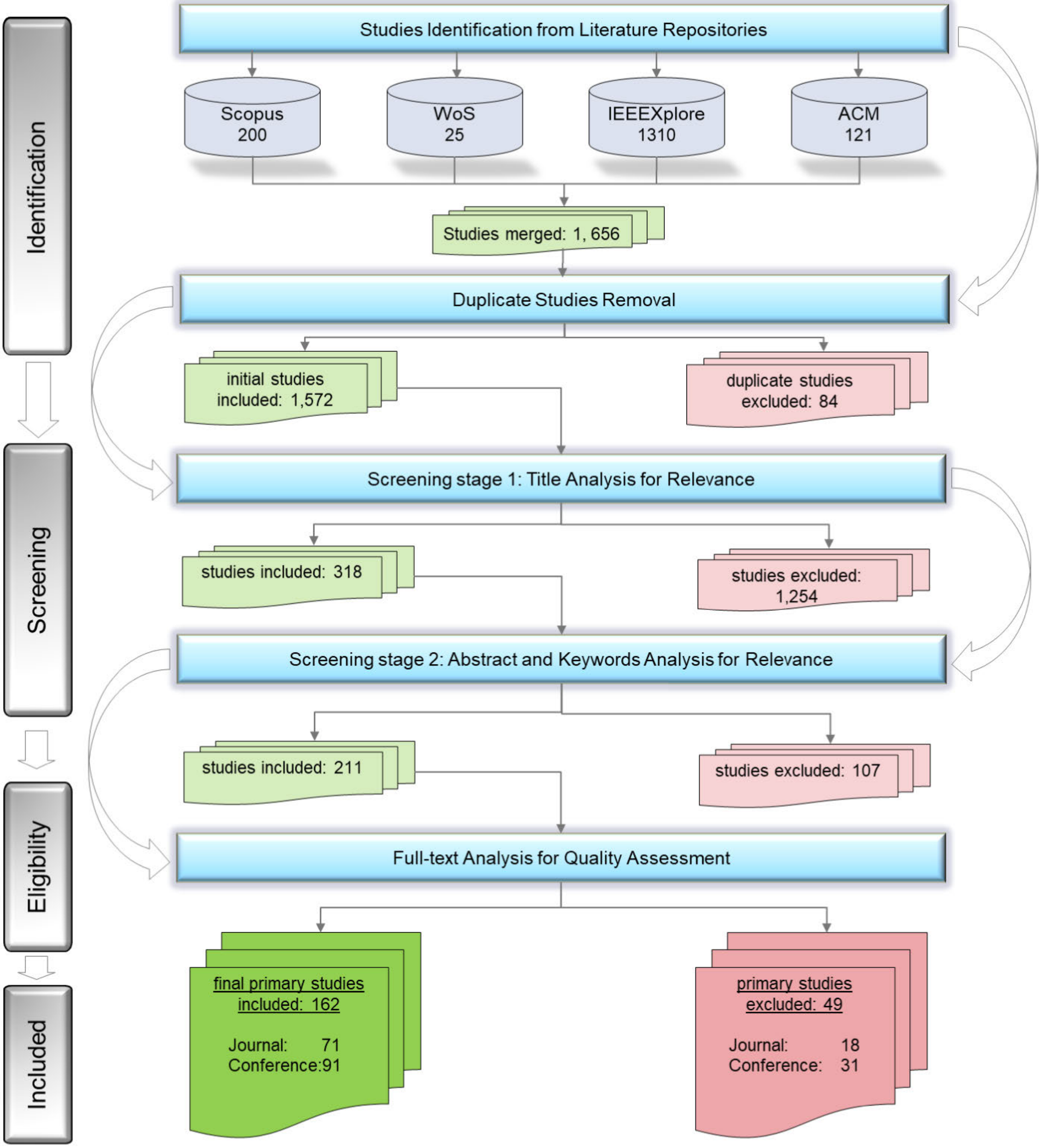
Identification
Screening
Eligibility
Included
Studies Identification from Literature Repositories
Scopus 200
WoS 25
IEEEXplore 1310
ACM 121
Studies merged: 1, 656
Duplicate Studies Removal
initial studies included: 1,572
duplicate studies excluded: 84
Screening stage 1: Title Analysis for Relevance
studies included: 318
studies excluded: 1,254
Screening stage 2: Abstract and Keywords Analysis for Relevance
studies included: 211
studies excluded: 107
Full-text Analysis for Quality Assessment
final primary studies included: 162
Journal: 71
Conference:91
primary studies excluded: 49
Journal: 18
Conference: 31

FIGURE 2. Primary studies selection process.

Screening Stage 2 – Abstract and keywords analysis: Doubtful studies were discussed by the team members in this stage. All the studies included in this stage were downloaded for the subsequent stage.

Screening Stage 3 – Full-text analysis: All the studies were available to all authors in Google Drive. This stage can also be considered the eligibility stage of the studies.

Table 4 lists the results of each stage concerning all databases.

Specifically, we excluded 1,254 irrelevant studies during stage I, and 107 in stage II. Then we assessed the actual relevance of the remaining potentially relevant 211 primary studies in stage III using the eligibility criteria described in Table 5. Study selection criteria for all three stages were the

TABLE 4. Overview of studies results in each step.

Source	Initial results	After title analysis	After abstract, keyword analysis	After quality assessment
Scopus	192	128	100	76
WoS	16	11	11	11
IEEE Xplore	1,247	168	92	70
ACM	117	11	8	5
Total	*1,572*	*318*	*211*	*162*

same. We recorded the reasons for the exclusion of studies not fulfilling the eligibility criteria.

Study Selection Criteria: Study selection or eligibility criteria affect the SLR results. It was defined during the protocol development. As the inclusion/exclusion criteria set the boundaries of the SLR and were based on our research question. We tested our criteria on a small number of primary studies first. Then we refined it before the actual search process started to save the search time.

The above criteria are the list of restrictions regarding the studies. These restrictions are by language, area of study, and primary studies. There were many irrelevant studies found that did not address any aspect of the research questions. Two researchers assessed each study and each disagreement was discussed and resolved.

E. STUDY QUALITY ASSESSMENT

A vital step in carrying out a systematic review is evaluating the quality of potential primary studies. This process is intended to evaluate the scientific credibility of studies and wipe out thematically irrelevant studies [19]. The studies were evaluated to answer our RQs to gauge the credibility of the research. The quality assessment questions for video captioning studies are indicated in Table 6 and the results of its analysis are presented in Table 7.

The studies were evaluated, and only those supporting our quality assessment questions were selected. The grades against the seven above-mentioned quality criteria items were formulated as follows:

- 1 for Yes
- 0.5 for Partially Yes
- 0 for No

Thus, a primary study was excluded if its mean grade was less than 5. Two authors performed this step and all discrepancies about the studies' quality were resolved mutually.

Table 7 shows that out of one hundred sixty-two studies, thirty-eight studies obtained a significant mean grade of 5, eighty-three studies gained a mean grade of 6, and forty-four studies earned a mean grade of 7. Although forty-nine studies were excluded due to insufficient mean grades. This table contains the studies year-wise. The top 9 studies were selected from the year 2023, 33 from the year 2022, 14 from the year 2021, 19 from the year 2020, 22 studies from the year 2019, 27 from the year 2018, 24 from the year 2017, 12 from the year 2016, 4 from the year 2015, and 1 from the year 2014 as depicted in Figure 3. Figure 6 depicts the proportion of selected primary studies distributed across different years, organized by databases.

Table 8 presents a comprehensive performance comparison of selected studies with the mean grades for each study measured as seven, across different evaluation metrics, including BLEU, METEOR, ROUGE, and CIDEr. Among the highlighted studies, S3 stands out with impressive scores across all metrics, particularly achieving a high CIDEr score of 119.5. S3 and S6 also demonstrate strong performance, scoring notably well in BLEU, METEOR, and ROUGE.

1) DATA EXTRACTION

Research questions were answered after examining the studies in the study selection step. In this step, the information required to answer our research question was taken from the selected primary papers. The data collected from 162 primary studies is displayed in Table 7 to answer the research question.

To avoid irrelevant data extraction, it was confirmed by the authors that the extracted data addressed the factors of primary question and secondary questions. Table 9 provides a comprehensive overview of the extracted data from selected studies, focusing on various key items relevant to the research questions. The table is designed to organize and present essential information for analysis and synthesis.

2) DATA SYNTHESIS

In this section, extracted data is presented for synthesis in tabular and graphical formats for information visualization. Table 10 and Table 11 show the list of the venues for the selected primary studies published along with some necessary details.

F. REVIEW REPORT

We have followed the PRISMA statement for reporting because it is a highly preferable reporting method for systematic reviews, and its results play a crucial role in the assessment of various research questions. This systematic review focuses on the evidence of the influence of deep learning methods compared to machine learning methods in the

TABLE 5. Inclusion/exclusion criteria for systematic review of video captioning studies.

Inclusion Criteria	Exclusion Criteria
1. Studies with full text available 2. Studies published in scientific journals or conference 3. Primary studies that proposed an approach for the video captioning process	1. Apart from English-language academic publications 2. Duplicate studies (by title/content) 3. Reviews, editorials, posters, short communications, and patents 4. Studies that do not validate the proposed methodology

TABLE 6. Quality criteria for a systematic review of video captioning studies.

1. Does the study clearly describe the aim of the research?
2. Does the study use a publicly available dataset?
3. Does the study specify the source of the dataset?
4. Does the study use relevant performance evaluation metrics for result analysis?
5. Were the results compared with other baseline studies?
6. Was the result comparison performed with the latest studies?
7. Does the conclusion relate to the aim of the research?

field of video captioning. The research question (RQ1) aimed to determine the superiority of deep learning methods over machine learning methods in the field of video captioning.

The results showed that the majority of selected studies in the field of video captioning have been conducted using deep learning methods, followed by a smaller number of studies that utilized machine learning methods. The usage statistics of these methods are presented in Table 13 and Table 14.

The second research question (RQ2) aimed to investigate which video captioning datasets are preferred for benchmarking purposes, relative to different parameter characteristics. The results revealed that the MSR-VTT dataset, in addition to MSVD, has been extensively used for video captioning. However, comparing the two datasets, Table 15 shows that the MSR-VTT dataset has a larger number of video clips, sentences, and vocabulary. However, according to Figure 7, the MSVD dataset outperformed the MSR-VTT dataset in evaluation metrics when different models were employed and were selected more frequently.

The third research question (RQ3) aimed to determine the most appropriate performance evaluation metric for studies of video captioning models, relative to parameter characteristics. The results showed that METEOR and BLEU were the most commonly used evaluation metrics in the selected studies, as shown in Figure 11. These findings provide insights into the preferred methods and metrics used in the field of video captioning and can serve as a reference for future research in this area.

II. RESULTS AND INTERPRETATION

The data extracted from selected primary studies are discussed and analyzed in this section. A total of 1,656 studies were retrieved and 162 papers based on study selection criteria (see Table 5), and assessment questions (see Table 6) were finalized for the analysis. Table 13 and Table 14 list the details of studies employing machine learning methods and deep learning techniques for feature extraction and caption generation methods respectively. Moreover, the research questions are answered in this section.

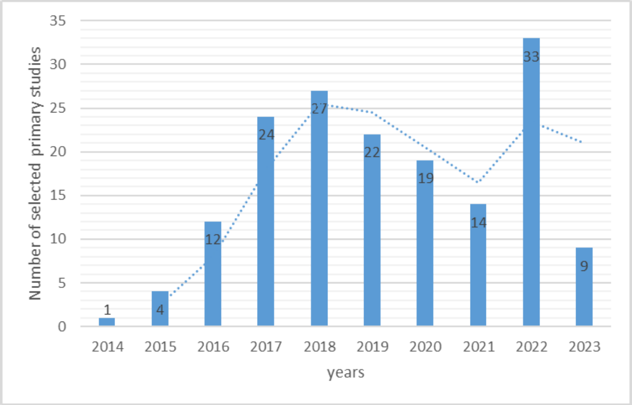

FIGURE 3. Yearly distribution of selected primary studies.

A. STUDY DEMOGRAPHIC

This subsection shows a general demographic overview of the study's data to show the overall productivity in the under-study research area. That's why we decided to extract the Publication venue and year distribution. The number of studies published in journals is higher than the conference proceedings.

Figure 3 shows that since 2017, the interest in video captioning has been significant (on average, twenty papers are published annually), with vastly higher increases in 2017–2022.

TABLE 7. Quality scores of selected studies.

	Scores							
Study ID	**Q#1**	**Q#2**	**Q#3**	**Q#4**	**Q#5**	**Q#6**	**Q#7**	**Sum**
S1 [20]	1	1	1	1	0.5	0.5	0	5
S2 [21]	1	1	1	0.5	1	0	1	5.5
S3 [22]	1	1	1	1	1	0	1	6
S4 [23]	1	1	1	1	1	0.5	1	6.5
S5 [24]	1	1	1	1	1	0.5	1	6.5
S6 [25]	1	1	1	1	1	1	1	7
S7 [26]	1	1	1	1	1	1	1	7
S8 [27]	1	1	1	1	1	1	1	7
S9 [28]	1	1	1	1	1	1	1	7
S10 [29]	1	0	1	1	1	0	1	5
S11 [30]	1	1	1	1	0	0	1	5
S12 [31]	1	1	1	1	1	0	0	5
S13 [32]	0.5	1	1	1	1	0	1	5.5
S14 [33]	1	1	0	1	1	1	0.5	5.5
S15 [34]	1	1	1	1	1	0	0.5	5.5
S16 [35]	1	1	0	1	1	0.5	1	5.5
S17 [36]	1	1	1	0.5	1	0	1	5.5
S18 [37]	1	1	1	1	1	1	1	7
S19 [38]	1	1	1	1	1	0	1	6
S20 [39]	1	1	1	1	1	0	1	6
S21 [40]	1	1	0	1	1	1	1	6
S22 [41]	1	1	1	1	1	0.5	1	6.5
S23 [42]	1	1	0	1	1	1	1	6
S24 [43]	1	1	1	1	1	0.5	1	6.5
S25 [44]	1	1	1	1	1	0.5	1	6.5
S26 [45]	1	1	1	1	1	0.5	1	6.5
S27 [46]	1	1	1	1	1	0.5	1	6.5
S28 [47]	1	1	1	1	1	0.5	1	6.5
S29 [48]	1	1	1	1	1	0.5	1	6.5
S30 [49]	1	1	1	1	1	1	0.5	6.5
S31 [50]	1	1	1	1	1	0.5	1	6.5
S32 [51]	1	1	1	1	1	0.5	1	6.5
S33 [52]	1	1	1	1	1	0.5	1	6.5
S34 [53]	1	1	1	1	1	1	0.5	6.5
S35 [54]	1	1	1	1	1	1	1	7
S36 [55]	1	1	1	1	0	0	1	5
S37 [56]	1	1	1	1	1	1	1	7
S38 [57]	1	1	1	1	1	1	1	7
S39 [58]	1	1	1	1	1	1	1	7
S40 [59]	1	1	1	1	1	1	1	7
S41 [60]	1	1	0	1	1	0	1	5
S42 [61]	0.5	1	1	1	1	0	0.5	5
S43 [62]	1	1	1	1	1	0	1	6
S44 [63]	1	1	1	1	1	0	1	6
S45 [64]	1	1	1	1	1	0	1	6
S46 [65]	1	1	1	1	1	0	1	6
S47 [66]	1	1	1	1	1	0	1	6
S48 [67]	1	1	0	1	1	1	1	6
S49 [68]	1	1	1	1	1	0.5	1	6.5
S50 [69]	1	1	1	1	1	0.5	1	6.5
S51 [70]	1	1	1	1	1	0.5	1	6.5
S52 [71]	1	1	1	1	1	1	1	7
S53 [72]	1	1	1	1	1	1	1	7
S54 [73]	1	1	1	1	1	1	1	7

TABLE 7. ***(Continued.)*** **Quality scores of selected studies.**

S55 [74]	1	1	1	1	0	0	1	5
S56 [75]	1	0	0	1	1	1	1	5
S57 [76]	1	1	1	0	1	0.5	1	5.5
S58 [77]	1	1	1	1	1	0	0.5	5.5
S59 [78]	1	1	1	1	1	0	1	6
S60 [79]	1	1	1	1	1	1	0	6
S61 [80]	1	1	1	1	1	0	1	6
S62 [81]	1	1	1	1	1	0	1	6
S63 [82]	1	1	1	1	1	0	1	6
S64 [83]	1	1	1	1	1	0	1	6
S65 [84]	1	1	1	1	1	0	1	6
S66 [85]	1	1	1	1	1	0.5	1	6.5
S67 [86]	1	1	1	1	1	0.5	1	6.5
S68 [87]	1	1	1	1	1	0.5	1	6.5
S69 [88]	1	1	1	1	1	1	0.5	6.5
S70 [89]	1	1	1	1	1	0.5	1	6.5
S71 [90]	1	1	1	1	1	1	0.5	6.5
S72 [91]	1	1	1	1	1	1	1	7
S73 [92]	1	1	1	1	1	1	1	7
S74 [93]	1	1	0	1	1	0	1	5
S75 [94]	1	1	0.5	0.5	1	0.5	0.5	5
S76 [95]	1	0	0	1	1	1	1	5
S77 [96]	1	1	1	1	1	0	0	5
S78 [97]	1	1	1	0.5	1	0	0.5	5
S79 [98]	1	1	1	1	1	0	0.5	5.5
S80 [99]	1	1	1	1	1	1	0	6
S81 [100]	1	1	1	1	1	0.5	0.5	6
S82 [101]	1	1	1	1	1	0	1	6
S83 [102]	1	1	1	1	1	0	1	6
S84 [103]	1	1	1	1	1	0	1	6
S85 [104]	1	1	1	1	1	0	1	6
S86 [105]	1	1	1	1	1	0	1	6
S87 [106]	1	1	1	1	1	0.5	1	6.5
S88 [107]	1	1	1	1	1	1	0.5	6.5
S89 [108]	1	1	1	1	1	0.5	1	6.5
S90 [109]	1	1	1	1	1	0.5	1	6.5
S91 [110]	1	1	1	1	1	0.5	1	6.5
S92 [111]	1	1	1	1	1	1	1	7
S93 [112]	1	1	1	1	1	1	1	7
S94 [113]	1	1	1	1	1	1	1	7
S95 [114]	1	1	1	1	1	1	1	7
S96 [115]	1	1	1	1	0	0	1	5
S97 [116]	1	1	1	1	1	0.5	0	5.5
S98 [117]	1	1	0	1	1	0.5	1	5.5
S99 [118]	1	1	0	1	1	0.5	1	5.5
S100 [119]	0.5	1	1	1	1	1	0.5	6
S101 [120]	1	1	1	1	1	0.5	0.5	6
S102 [121]	1	0	1	1	1	1	1	6
S103 [122]	1	1	1	1	1	0	1	6
S104 [123]	1	1	1	1	1	0	1	6
S105 [124]	1	1	1	1	1	1	0.5	6.5
S106 [125]	1	1	1	1	1	0.5	1	6.5
S107 [126]	1	1	1	1	1	0.5	1	6.5
S108 [127]	1	1	1	1	1	0.5	1	6.5
S109 [128]	1	1	1	1	1	1	0.5	6.5
S110 [129]	1	1	1	1	1	1	1	7
S111 [130]	1	1	1	1	1	1	1	7

TABLE 7. ***(Continued.)*** **Quality scores of selected studies.**

S112 [131]	1	1	1	1	1	1	1	7
S113 [132]	1	1	1	1	1	1	1	7
S114 [133]	1	1	1	1	1	1	1	7
S115 [134]	1	1	1	1	1	1	1	7
S116 [135]	1	1	1	1	1	1	1	7
S117 [136]	1	1	1	1	1	1	1	7
S118 [137]	1	1	1	1	0	0	1	5
S119 [138]	1	1	1	1	1	1	1	7
S120 [139]	1	1	1	1	1	1	1	7
S121 [140]	1	1	1	1	1	1	1	7
S122 [141]	1	1	1	1	1	1	1	7
S123 [142]	1	0.5	0	1	1	0.5	1	5
S124 [143]	1	0.5	0.5	1	1	1	0	5
S125 [144]	1	1	1	1	1	0.5	0	5.5
S126 [145]	1	1	1	1	1	1	0	6
S127 [146]	1	1	1	1	1	1	0	6
S128 [147]	1	1	1	1	1	0.5	1	6.5
S129 [148]	1	1	1	1	1	0.5	1	6.5
S130 [149]	1	1	1	1	1	1	0.5	6.5
S131 [150]	1	1	1	1	1	1	0.5	6.5
S132 [151]	1	1	1	1	1	1	0.5	6.5
S133 [152]	1	1	1	1	1	0.5	1	6.5
S134 [153]	1	1	1	1	1	1	0.5	6.5
S135 [154]	1	1	1	1	1	1	0.5	6.5
S136 [155]	1	1	1	1	1	1	0.5	6.5
S137 [156]	1	1	1	1	1	1	1	7
S138 [157]	1	1	1	1	1	1	1	7
S139 [158]	1	1	1	1	1	1	1	7
S140 [159]	1	1	1	1	1	1	1	7
S141 [160]	1	1	1	1	1	1	1	7
S142 [161]	1	1	1	1	1	1	1	7
S143 [162]	1	1	1	1	1	1	1	7
S144 [163]	1	1	1	1	1	1	1	7
S145 [164]	1	1	1	1	1	1	1	7
S146 [165]	0.5	1	1	1	1	0	0.5	5
S147 [166]	1	1	1	1	0	0	1	5
S148 [167]	0.5	1	0.5	1	1	1	0.5	5.5
S149 [168]	0.5	1	1	1	1	1	0	5.5
S150 [169]	0.5	1	1	1	1	1	0	5.5
S151 [170]	1	1	0	1	1	1	0.5	5.5
S152 [171]	1	1	1	1	1	1	0	6
S153 [172]	1	1	1	1	1	1	0	6
S154 [173]	0.5	1	1	1	1	1	0.5	6
S155 [174]	1	1	1	1	1	1	0	6
S156 [175]	1	1	1	1	1	0.5	1	6.5
S157 [176]	1	1	1	1	1	1	1	7
S158 [177]	0.5	1	1	1	1	1	0.5	6
S159 [178]	1	1	1	1	1	1	1	7
S160 [12]	1	1	1	1	1	1	1	7
S161 [179]	1	1	1	1	1	1	1	7
S162 [10]	1	0.5	0	1	1	1	1	5.5

TABLE 8. Performance comparison of highest mean grade studies.

Study ID	BLEU-4	METEOR	ROUGE	CIDEr
S6 [25]	**<u>59.7</u>**	37.3	74.3	101.5
S7 [26]	1.91	10.24	-	32.83
S8 [27]	0.392	0.332	0.435	-
S9 [28]	55.0	36.6	73.9	100.6
S18 [37]	12.44	17.55	32.91	28.32
S35 [54]	59.8	38.5	88.2	97.8
S37 [56]	42.0	28.8	62.0	54.2
S38 [57]	67.4	**<u>41.7</u>**	79.2	**<u>119.5</u>**
S39 [58]	1.68	7.91	-	23.02
S40 [59]	41.8	30.8	65.7	74.4
S52 [71]	54.3	34.0	**<u>80.3</u>**	-
S53 [72]	32.6	-	57.3	51.2
S54 [73]	45.3	32.6	69.4	75.8
S72 [91]	54.1	35.6	-	79.6
S73 [92]	48.8	33.4	69.7	82.0
S92 [111]	40.4	28.1	60.7	47.1
S93 [112]	56.9	36.2	-	90.6
S94 [113]	47.9	35.0	71.5	78.1
S95 [114]	0.93	8.82	-	30.68
S110 [129]	39.8	26.5	-	41.1
S111 [130]	2.30	9.65	19.29	12.68
S112 [131]	53.0	34.7	65.9	79.4
S113 [132]	42.5	32.0	68.8	59.0
S114 [133]	1.62	10.33	-	25.24
S115 [134]	52.3	33.3	69.6	76.5
S116 [135]	52.02	32.18	-	-
S117 [136]	8.45	14.75	25.88	-
S119 [138]	50.0	-	-	96.3
S120 [139]	52.3	34.1	69.8	80.3
S121 [140]	54.57	34.61	-	83.85
S122 [141]	50.0	-	-	94.3
S137 [156]	51.1	33.6	-	74.8
S138 [157]	44.3	32.1	68.9	68.4
S139 [158]	-	0.084	0.229	0.084
S140 [159]	0.458	0.333	0.697	0.730
S141 [160]	48.76	34.36	-	80.45
S142 [161]	42.5	31.0	-	-
S143 [162]	52.8	33.5	-	74.0
S144 [163]	0.511	0.327	-	0.675
S145 [164]	44.2	29.4	62.6	50.5
S157 [176]	0.499	0.326	-	0.658
S159 [178]	0.4192	0.2960	-	0.5167
S160 [12]	-	29.8	-	-
S161 [179]	35.09	29.26	-	-

As seen in Figure 4, conferences were the most popular venues for publishing the selected primary studies and the most frequently used venue was the IEEE Conference on Computer Vision and Pattern Recognition known as CVPR (see Table 11).

Figure 3 presents the number of primary studies per year and an increase in the under-study research area. Most of the video captioning studies were presented at conferences as shown in Figure 4. Figure 5 illustrates the distribution of primary studies categorized by databases.

Table 12 lists the most influential studies so far in the area of video captioning. Below is a short methodological analysis of these studies:

S162 proposed a Seq2Seq model for video captioning that uses a convolutional neural network (CNN) as the encoder and a recurrent neural network (RNN) as the decoder. The

TABLE 9. Data extraction from selected studies.

Extracted Data Items	Target
1. Publication year	Overview
2. Publication type	Overview
3. Deep learning methods	RQ1
4. Machine learning methods	RQ1
5. Methods characteristics	RQ1
6. Video captioning datasets	RQ2
7. Dataset characteristics	RQ2
8. Performance evaluation metrics	RQ3
9. Metric characteristics	RQ3

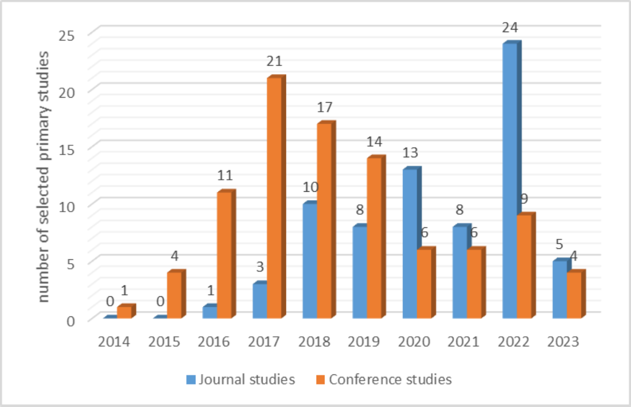

FIGURE 4. Document type-split of yearly distribution of selected primary studies.

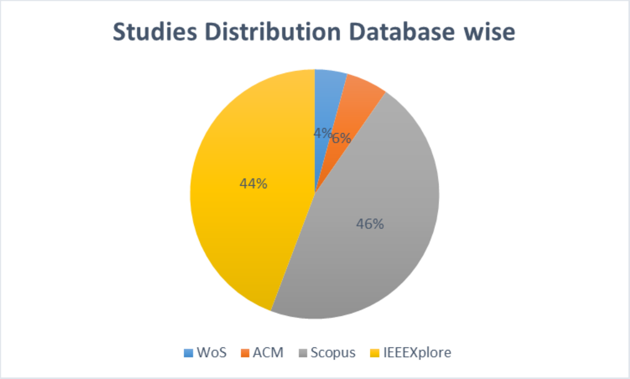

FIGURE 5. Database-wise primary studies distribution.

encoder is a convolutional neural network (CNN) that processes video frames and encodes them into a fixed-length representation. The decoder is a recurrent neural network (RNN) that generates the video caption word by word. S161 proposes a novel two-stage approach for video captioning that leverages the temporal structure of videos to generate high-quality captions. The main contribution of the paper is a novel approach to video caption generation that leverages the inherent temporal relationships between video frames and captions. The authors propose a two-stage approach for video captioning that first generates a set of candidate captions and then refines these captions to produce the final captions. In the first stage, the model generates candidate captions by predicting the next word in the caption based on the current state of the video, using a recurrent neural network (RNN) with an attention mechanism. In the second stage, the model refines the candidate captions by considering the video content and the relationship between video frames and captions.

S160 presents a novel approach for video captioning that generates longer and more semantically meaningful captions by leveraging a hierarchical recurrent neural network and an attention mechanism. S158 proposes a video captioning model that jointly models the embeddings of video frames and the translations between video frames and captions. The model consists of two components: an encoder that encodes video frames into continuous embeddings, and a decoder that generates captions from the embeddings.

S142 makes a significant contribution to the field of video captioning, by combining an attention-based LSTM network with a semantic consistency loss and demonstrating its effectiveness in generating coherent and semantically consistent captions for video sequences.

S57 proposes a video representation model based on HRNNs, which consist of multiple levels of RNNs that process different levels of information in the video. S146 proposes a video captioning model that first transfers the semantic attributes from a pre-trained attribute recognition network to the video frames, and then uses a sequence-to-sequence (Seq2Seq) model to generate captions based on the video frames and their associated semantic attributes.

S127 makes a valuable contribution to the field of video captioning, by proposing a novel video captioning model based on attention-based multimodal fusion and demonstrating its effectiveness in generating captions that are relevant to both the video frames and the audio track. S68 proposes a new video captioning model with a Spatial-Temporal Attention Mechanism (STAT) to address the limitations of previous video captioning models, which often struggle to effectively capture the relationships between video frames and captions. They argue that these models can be improved by using attention mechanisms, which can weigh the importance of different video frames when generating captions.

They use a combination of two attention modules: a spatial attention module that weighs the importance of different parts of the video frames, and a temporal attention module that weighs the importance of different frames in the video. S122 incorporated a high-level policy network that generates a sequence of coarse-grained action descriptions, and a low-level value network that refines these descriptions into captions. Strengths of the proposed approach include its ability to effectively model the relationships between the video and language modalities, and its flexibility in handling diverse video content and captions.

The above papers represent some of the most influential primary studies in the field of video captioning and have significantly advanced the state-of-the-art in this area.

TABLE 10. Selected journals.

Journal Name	Number of studies	Published by	Impact Factor	Quarter Category
ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)	2	ACM	3.144	Q1
Applied Sciences (Switzerland)	1	MDPI	2.679	Q2
China Communications	1	China Communications Magazine	2.688	Q2
Computer Vision and Image Understanding	1	Academic Press Inc.	5.53	Q2
Electronics (Switzerland)	1	MDPI	2.657	Q2
Expert Systems with Applications	1	Elsevier	8.665	Q1
IEEE Access	5	IEEE	3.367	Q1
IEEE Internet of Things Journal	1	IEEE	9.936	Q1
IEEE MultiMedia	1	IEEE	3.556	Q1
IEEE Robotics and Automation Letters	1	IEEE	3.741	Q2
IEEE Transactions on Artificial Intelligence	1	IEEE	7.25	Q1
IEEE Transactions on Circuits and Systems for Video Technology	8	IEEE	4.685	Q1
IEEE Transactions on Image Processing	7	IEEE Signal Processing Society	10.86	Q1
IEEE Transactions on Multimedia	7	IEEE	5.452	Q1
IEEE Transactions on Neural Networks and Learning Systems	2	IEEE	14.26	Q1
IEEE Transactions on Pattern Analysis and Machine Intelligence	2	IEEE	24.31	Q1
International Journal of Advanced Science and Technology	1	Science and Engineering Research Support Society	0.108	Q2
International Journal of Electrical and Computer Engineering	1	Institute of Advanced Engineering and Science (IAES)	1.616	N/A
International Journal of Information Technology (Singapore)	1	World Scientific	1.406	Q2
International Journal of Intelligent Systems	1	Hindawi	8.709	Q1
Journal of Big Data	2	Springer	4.426	Q1
Journal of China Universities of Posts and Telecommunications	1	Beijing University of Posts and Telecommunications	0.27	Q4
Journal of Electronic Imaging	1	IS&T, SPIE	0.945	Q4
Journal of Information Processing Systems	1	KIPS	1.518	N/A
Journal of King Saud University - Computer and Information Sciences	1	Elsevier Science Inc.	6.05	Q1
Journal of Visual Communication and Image Representation	2	Elsevier Science Inc.	2.259	Q3
Mathematical Problems in Engineering	1	Hindawi	1.58	Q4
Multimedia Tools and Applications	1	Springer Netherlands	2.757	Q2
Neural Computing and Applications	1	Springer London	5.606	Q1
Neurocomputing	3	Elsevier Science Inc.	5.719	Q2
Pattern Analysis and Applications	1	Springer London	2.764	Q3
Pattern Recognition Letters	1	Elsevier, International Association for Pattern Recognition	2.81	Q2
PeerJ Computer Science	1	San Francisco CA: PeerJ Inc.	2.41	Q2
Quarterly of Applied Mathematics	1	American Mathematical Society	0.815	Q3
Sensors	2	MDPI	3.576	Q2
Sustainability (Switzerland)	1	MDPI AG	4.17	Q2
Traitement du Signal	1	Lavoisier	2.64	Q3
Visual Computer	1	Springer	2.601	Q2
World Wide Web	1	Springer	2.716	Q2

TABLE 11. Selected conferences.

Conference Name	Number of studies	H-Index
IEEE International Conference on Image Processing (ICIP)	1	108
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	1	174
International Conference on Smart Systems and Inventive Technology (ICSSIT)	1	10
International Joint Conference on Neural Networks (IJCNN)	2	90
IEEE Workshop on Applications of Computer Vision (WACV)	2	22
International Conference for Emerging Technology (INCET)	1	43
International Conference on Advanced Computing and Communication Systems (ICACCS)	1	14
International Conference on Artificial Intelligence and Big Data (ICAIBD)	1	10
Proceedings of the ACM Turing Celebration Conference – China	1	-
ACM International Conference on Multimedia	4	71
ACM International Conference Proceeding Series	1	128
ACM Multimedia Conference	7	-
British Machine Vision Conference	1	75
Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies	1	105
Conference on Artificial Intelligence	1	180
European Conference on Computer Vision (ECCV)	2	186
IEEE Automatic Speech Recognition and Understanding Workshop	1	27
IEEE Computer Society Conference on Computer Vision and Pattern Recognition	3	408
IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	21	104
IEEE Globecom Workshops	1	29
IEEE International Conference on Advanced Robotics and Mechatronics	1	5
IEEE International Conference on Computer Vision	3	280
IEEE International Conference on Engineering, Technology and Innovation	1	11
IJCAI International Joint Conference on Artificial Intelligence	1	142
International Conference on Computational Linguistics	2	62
International Conference on Computer and Communication Engineering Technology	1	5
International Conference on Computer Vision Workshop	1	39
International Conference on Image, Vision and Computing	1	14
International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)	1	19
International Conference on Language Resources and Evaluation	1	19
International Conference on Pattern Recognition	2	113
International Joint Conference on Artificial Intelligence	1	142
Lecture Notes in Computer Science	3	415
Proceedings of SPIE - The International Society for Optical Engineering	1	187
Proceedings of the International Joint Conference on Neural Networks	1	82
Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition	2	470
TREC Video Retrieval Evaluation	3	12
Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)	1	15
Big Data, Cloud Computing, Data Science & Engineering (BCD)	1	4
IEEE Information Technology, Networking, Electronic and Automation Control Conference	1	10
IEEE International Conference on Big Knowledge (ICBK)	1	8
IEEE International Conference on Multimedia Big Data (BigMM)	2	10
IEEE Workshop on Applications of Computer Vision (WACV)	2	22
International Conference on Automation and Computing (ICAC)	1	12
International Conference on Computational Science/Intelligence and Applied Informatics (CSII)	1	3
International Conference on Computing and Networking Technology (ICCNT)	1	7
International Conference on Computing Communication and Networking Technologies (ICCCNT)	1	17
International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)	1	8
International Image Processing, Applications and Systems Conference (IPAS)	1	8

TABLE 12. List of most influential primary studies.

Study ID	Title	# of citations
S162	Sequence to Sequence -- Video to Text [12]	1476
S161	Describing videos by exploiting temporal structure [178]	1143
S160	Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks [174]	627
S158	Jointly Modeling Embedding and Translation to Bridge Video and Language [175]	598
S142	Video Captioning With Attention-Based LSTM and Semantic Consistency [144]	521
S157	Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning [176]	434
S146	Video Captioning with Transferred Semantic Attributes [162]	337
S127	Attention-Based Multimodal Fusion for Video Description [147]	324
S68	STAT: Spatial-Temporal Attention Mechanism for Video Captioning [80]	256
S122	Video Captioning via Hierarchical Reinforcement Learning [128]	237

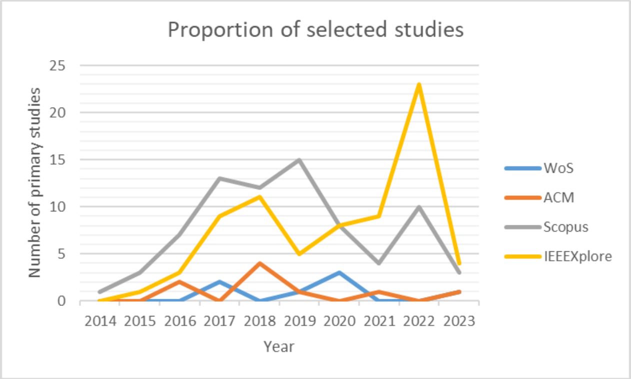

FIGURE 6. Year-wise database proportion of selected primary studies.

B. RESULTS OF RQ1: WHAT EVIDENCE INDICATES THAT DEEP LEARNING METHODS ARE MORE COMMONLY UTILIZED THAN MACHINE LEARNING METHODS IN VIDEO CAPTIONING?

There is a plethora of numerous algorithms and approaches available for the video captioning job; however, they may be categorized essentially into two primary groups; the template-based language models and the sequence learning models.

The traditional template-based model approach breaks the sentence up into many parts and determines the specific grammar rule in advance (e.g., Subject, Verb, Object). Several studies employ these sentence fragments to line up each component with words that were derived from visual imagery using object recognition, and they then create a phrase with linguistic constraints. A translatable mapping between the information of a video and a sentence is directly learned using the deep learning approach, which makes use of sequence learning models. Each category works in a two-stage process; feature extraction, and caption generation.

There are various aspects to consider when selecting a video feature extraction method. The type of features is the most important aspect as various types of features can be extracted from video, including low-level features such as color, texture, and motion, and high-level features such as objects, scenes, and actions.

Representation also plays an important role because extracted features can be represented in different ways, such as pixel-based, region-based, and object-based. The pixel-based representation uses the raw pixel values of a video frame, while region-based and object-based representations use regions or objects, respectively, in a video frame. Spatial and temporal resolution is also important as it determines the detail and accuracy of the extracted features. Video feature extraction methods can operate on different spatial and temporal resolutions. For example, some methods can operate on a single frame, while others can operate on multiple frames to capture temporal information. Scale affects the ability of the method to handle videos of different sizes and resolutions. Some video feature extraction methods can operate at different scales, such as multi-scale or scale-invariant, to capture features at different levels of abstraction. Computational complexity determines the processing speed and efficiency of the method. Different video feature extraction methods can have different computational complexities, which may impact their real-time applicability. The robustness of a video feature extraction method refers to its ability to work well under various conditions, such as illumination changes, camera motion, and background clutter. Different video feature extraction methods can have different compatibility with different video analysis tasks, such as object recognition, scene classification, and action recognition. Thus, selecting a video feature extraction method requires balancing these different aspects to achieve the desired results. To address the first RQ, proposed methods of selected studies were analyzed under these stages, shown in Table 13, and Table 14.

In Table 13 and Table 14, the terms "high" and "low" in the context of computational complexity refer to the computational expenses associated with each model. When describing computational complexity as "high," it indicates that the corresponding model demands substantial computational resources, encompassing significant processing power, memory utilization, or time for execution. On the other hand,

TABLE 13. Applied models and their aspects in the selected studies for feature extraction.

Model name	# of times adopted	Features extraction	Representation	Spatial resolution	Temporal resolution	Scale	Computational complexity	Robustness	Compatibility
Deep learning models									
2D-CNN	57	low-level	pixel-based	single-frame	×	Multiple	high	✓	✓
C3D-Net	44	low-level	pixel-based	×	multiple-frames	multiple	high	✓	✓
R-CNN	2	high-level	region-based	single-frame	×	multiple	high	✓	✓
Mask R-CNN	1	high-level	region-based	single-frame	×	multiple	high	✓	✓
Faster R-CNN	1	high-level	region-based	single-frame	×	multiple	high	✓	✓
GCN (Gated Convolutional Network)	1	high-level	region-based	×	multiple-frames	multiple	high	✓	✓
ResNet-50	3	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
ResNet-101	2	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
ResNet-152	6	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
ResNeXt-101	5	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
VGG-16	13	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
VGG-19	2	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
VGGish-Net	1	low-level	audio-based	×	✓	single	low	✓	✓
EfficientNet	1	high-level	pixel-based	single-frame	×	multiple	low	✓	✓
Inception-V3	3	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
Inception-ResNet-v2	2	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
NAS-Net	1	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
Tiny-YOLO	1	high-level	pixel-based	single-frame	×	single	high	✓	✓
GoogLeNet	5	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
IC3D-Net	1	high-level	pixel-based	×	multiple-frame	multiple	high	✓	✓
Transformer	2	high-level	pixel-based, learned embeddings	×	multiple-frame	multiple	high	✓	✓
ViT	1	high-level	pixel-based	single-frame	×	multiple	high	✓	✓
Attention mechanism	8	context-aware	pixel-based, audio-based	single-frame	✓	multiple	high	✓	✓
LSTM	22	high-level	pixel-based, audio-based	×	operates on sequences	single	high	✓	✓
Deep LSTM	1	high-level	pixel-based, audio-based	×	operates on sequences	single	high	✓	✓

TABLE 13. *(Continued.)* Applied models and their aspects in the selected studies for feature extraction.

Bi-LSTM	1	sequential	feature-based	✓	✓	multiple	high	✓	✓
GRU	2	sequential	feature-based	✓	✓	single	low	✓	✓
Machine learning models									
(CRF) Conditional Random Field	1	high-level	pixel-based	single-frame	✓	multiple	high	✓	✓
TvL1	1	optical flow	optical flow	×	✓	multiple	high	✓	✓
MFCC	2	audio	audio-based	×	✓	single	high	✓	✓
GAN	1	synthetic	pixel-based	✓	✓	multiple	high	depends on the quality of generated data	✓

TABLE 14. Applied models and their aspects in the selected studies for caption generation.

Model name	# of times adopted	Feature Extraction	Decoding Strategy	Pretraining	Training Data	Regularization	Data Augmentation	Computational Expense	Challenges
Deep Learning									
RNN	7	Yes	Greedy/Beam Search/Sampling	Optional	Large, Diverse	Dropout/Weight Decay	Optional	High	Vanishing Gradients
LSTM	96	Yes	Greedy/Beam Search/Sampling	Optional	Large, Diverse	Dropout/Weight Decay	Optional	High	Vanishing Gradients
GRU	12	Yes	Greedy/Beam Search/Sampling	Optional	Large, Diverse	Dropout/Weight Decay	Optional	High	Vanishing Gradients
Attention mechanism	35	Yes	Attention-Based	No	Large, Diverse	Dropout/Weight Decay	Optional	High	Attention-based biases
Temporal Attention Mechanism	1	Yes	Attention-Based	No	Large, Diverse	Dropout/Weight Decay	Optional	High	Attention-based biases
Transformer	8	Yes	Attention-Based	No	Large, Diverse	Dropout/Weight Decay	Optional	High	Attention-based biases
MLP	1	Yes	Greedy/Beam Search/Sampling	No	Large, Diverse	Dropout/Weight Decay	Optional	Low	Limited Representational Power
Machine Learning									
(SMT) Statistical Machine Translation	1	No	Rule-Based	No	Large, Diverse	No	No	Low	Limited Representational Power
CRF	1	No	Rule-Based	No	Large, Diverse	No	No	Low	Limited Representational Power

when denoted as "low," it signifies that the model exhibits minimal computational requirements, highlighting efficiency in terms of processing power, memory usage, or execution time. These terms are used as qualitative descriptors to succinctly convey the computational demands of the various models considered in the study.

We can conclude from the above table that three major categories of algorithms that recently shown impressive performance in feature extraction. C3D-Net is a neural architecture made up of several 3D convolutional layers that follow one another. Its training has the drawback of needing an excessive number of instances with labels. Most of the selected studies adopted its 2D and 3D architecture for feature extraction in video captioning. Along with these two, many variants of CNN have been implemented successfully. ResNet comes in a variety of flavors to address the vanishing gradient issue. For instance, training a ResNet-152 involves a lot of computations because it has roughly 60M parameters. This increases the amount of training time and energy needed. In comparison to ResNet-152, VGGNet not only has more parameters and FLOP but is also less accurate. Some selected studies deployed spatial exploitation-based CNN architectures e.g., VGG, and GoogLeNet which proved to be computationally costly due to the fully linked layers. A VGGNet with lower accuracy, for example, requires more time to train. Some other selected studies have used shortcut paths to give users the ability to skip certain layers, for example, in ResNet, a multipath CNN architecture. This improves the ability of the signals to be easily transmitted in both forward and backward directions. However, it brought up the problem of having to re-learn redundant feature maps. ResNeXt; a width-based variant of CNN that supports the concept of layer widening; is adopted by many selected studies that increased the cardinality to offer different transformations at each layer, although this results in a large computational cost. Many selected studies adopt the hybrid model to extract the features from videos and utilize the fusion of transformer or RNN variants with CNN. Most of the hybrid models for feature extraction adopted the diversification of CNN with LSTM and attention mechanism. For the sub-task of video feature extraction, a total of 189 times deep learning methods were employed in the selected studies, and in contrast, only five machine learning methods were used in the selected studies which mark a vital difference.

Different aspects of language models have an impact on their performance and efficiency. Feature extraction is one such aspect that determines the type and quality of features used to describe the input data. The decoding strategy refers to the method used to generate captions from the extracted features. Pretraining is a technique that utilizes a pre-trained model on a large dataset to initialize the parameters of a language model for the caption generation task. Training data is also an important aspect as it determines the quality of the captions generated by the model. Regularization is a technique used to prevent overfitting in the model, while data augmentation is used to increase the size of the training data. Computational expense is another important aspect as the model needs to be efficient in terms of computation time and memory usage. Challenges in the task of caption generation include dealing with diverse and complex inputs, handling large amounts of data, and generating captions that are accurate and coherent. Overall, a balanced consideration of these aspects can lead to better performance and efficiency of language models for the task of caption generation. Table 14 shows a comprehensive overview of these aspects.

Table 14 shows that RNN and LSTM have been adopted mostly for caption generation tasks in the selected studies of video captioning. A more advanced technique adopted for caption production is the attention mechanism.

In video captioning, deep learning methods have been more influential than traditional machine learning methods in recent years. This is due to the ability of deep neural networks to capture complex patterns and relationships in large amounts of data. Deep learning methods were adopted 160 times, and only 2 machine learning methods were adopted for the caption generation subtask. For example, in video captioning tasks, deep learning models such as encoder-decoder architectures and attention-based models have been shown to perform well in generating captions that are both descriptive and semantically accurate. These models have been trained on large datasets and use techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to model both spatial and temporal relationships in video frames and captions. On the other hand, traditional machine learning methods, such as support vector machines (SVMs) and decision trees, may struggle to capture the complex relationships between video frames and captions and produce captions that are less accurate and descriptive.

So, the choice of a video feature extraction method depends on the specific requirements of a video analysis task, including the type of features to be extracted, the representation of the features, the spatial and temporal resolution, the scale of the features, the computational complexity, the robustness, and the compatibility with the analysis task. However, it is important to note that deep learning methods are not always the best solution for every problem, and the choice of method depends on the specific task and available data. In some cases, traditional machine learning methods may still perform well and be a more suitable choice.

C. RESULTS OF RQ2: WICH VIDEO CAPTIONING DATASETS RELATIVE TO DIFFERENT PARAMETER CHARACTERISTICS ARE USED PREFERABLY FOR THE BENCHMARKING?

The use of contemporary video captioning methods has become more popular recently. However, the success of those techniques depends on the availability of datasets. Early attempts focused on building datasets for straightforward situations and closed domains. Later work improved the data collection and scalability by utilizing the internet and movie description services.

The size of a video captioning dataset refers to the number of video-caption pairs included. The video's diversity refers to the diversity of the video content, such as topics, styles, and camera views. The caption's quality is another important characteristic, as the captions should be well-written, grammatically correct, and accurately describe the video content.

Annotated information includes additional information such as labels, keywords, and emotions that can be used to further enhance the captioning models. Temporal information is information about the timing of events in the video, which is critical for generating accurate captions. Multi-modality refers to the ability to process different modalities such as audio, text, and visual information. Evaluation protocol is how the performance of a video captioning model is measured and compared to other models. These characteristics of video captioning datasets greatly influence the quality and performance of video captioning models. To address the 2nd RQ, the key aspects of the datasets employed frequently in

TABLE 15. List of selected video captioning datasets.

Dataset, Year	Domain	Videos	Clips	Standard split of clips	Sentence source	Sentences	Vocabulary size	clip-to-sentence ratio	Avg duration /video	employed in selected studies
MSVD, 2011 [180]	YouTube	1,970	1,970	training: 1,200 validation: 100 testing: 670	AMT workers	70,028	16,000	40	10s	96
MSR-VTT, 2016 [181]	Open	7,180	10,000	training: 6,513 validation: 497 testing: 2,990	AMT workers	200,000	29,000	20	20s	75
M-VAD, 2019 [182]	Movie	92	49,000	training: 19,023 validation: 2,976 testing: 2,836	DVS	55,904	18,000	1	6s	17
MPII-MD, 2015 [183]	Movie	94	68,337	training: 83 validation: 8 testing: 14	Script+ DVS	68,375	25,000	1	4s	12
ActivityNet, 2017 [148]	Open	20,000	20,000	training: 10,024 validation: 4,926 testing: 5,044	ActivityNet Annotators	100,000	10,000	5	180s	10
Charades, 2016 [184]	Indoor	157	9,848	train: 7,986 validation: 1,863 test: 2,000	AMT workers	27,847	27,847	1.5	30s	6

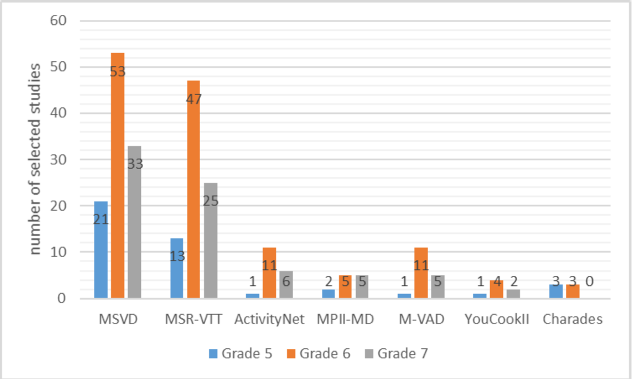

FIGURE 7. Quality grades distribution of popular datasets.

the selected studies are summarized here in Table 15 which could have dynamic effects on the performance of the method proposed.

Utilizing various datasets always affects a model's performance differently. That's why authors have selected different models and datasets with different evaluation metrics. Popular datasets like MSVD, MSR-VTT, M-VAD, MPII-MD, ActivityNet, and Charades were utilized in the selected studies. Each of the video captioning datasets listed in Table 16 has unique characteristics that make it influential in different ways.

MSVD (Microsoft Video Description), being one of the first and most widely used datasets, is the most popular and performs the best among all video captioning models. To develop the MSVD, the Mechanical Turk employees got paid for watching short video clips, describing their contents, and gathering 120K sentences over the summer of 2010. To prevent bias from lexical choices in the descriptions, the audio has been muted in all of the clips. More than 2,000 video clips are described in a collection of nearly parallel entries as a result. Due to the workers' encouragement to perform the activity in the language of their choosing, the data includes both bilingual and paraphrase alternations. This strength of the dataset makes it appealing to incorporate in as many experiments as possible, showing 28 studies with grade 5, 49 studies with grade 6, and 19 studies with grade 7 in Figure 7. A total of 96 studies have utilized the MSVD dataset. Its simplicity and early use make it influential, but its relatively small size and lack of temporal information make it challenging to use for training and evaluating video captioning models.

MSR-VTT (Microsoft Research Video to Text) is the second most frequently used dataset in quality studies. It, as a large-scale dataset, was introduced in 2016 to overcome the simplicity of current benchmark datasets with limited videos and simple descriptions of specific domains. This dataset is for the open-source video captioning domain that consists of 10k video clips from 20 categories and Amazon Mechanical Turks have labeled with 20 English sentences for each clip. 75 is a great score for this dataset's quality studies usage shown in Table 15. Its size and diversity make it a popular choice for training and evaluating video captioning models.

M-VAD (Montreal Video Annotation Dataset); a new dataset in the movie domain was developed in 2019 to address the lack of character-specific visual annotations in the existing movie description datasets. The M-VAD database includes annotations for characters' visual appearances with the face bounding boxes and, when accessible, associations

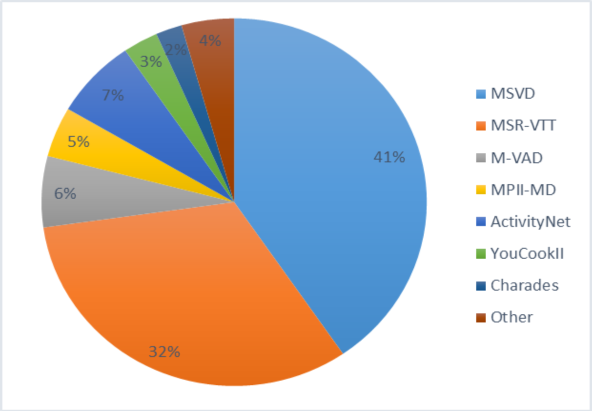

FIGURE 8. **Datasets usage frequency in selected studies.**

with their textual mentions. A semi-automatic approach was used to identify and annotate the participant's features in every movie's video clip. Its annotated information and multi-modality make it a useful dataset for evaluating the ability of video captioning models to generate captions that describe not only what is happening in the video, but also how it is happening.

MPII-MD (Max-Planck-Institut für Informatik-Movie Description) was established in 2015 to give transcribed audio descriptions of movies with the idea that audio descriptions (ADs) are significantly visual and descriptive to the scripts. One sentence from the movie's script and the audio caption information serve as the link between each clip. To ensure that the audio and visual information was consistent, each sentence was manually connected to the related video clip. Its size and the high quality of the captions make it a useful dataset for training and evaluating video captioning models.

ActivityNet Captions, which was introduced in 2017, revolutionized the field of video captioning by offering the ability to recognize each event in a video and describe it in natural language. The dense-captioning events method, which involves identifying and describing events in a video, is demonstrated in this groundbreaking dataset. Its size and diversity make it a useful dataset for training and evaluating video captioning models.

The Charades dataset, released in 2016, shows everyday indoor activities and lasts, on average, 30 seconds. It is a large-scale video captioning dataset that contains captions for complex activities in the videos, such as cooking, cleaning, and playing games. Each movie in this dataset has a variety of free-text annotations, action labels, action intervals, and classes of interacting objects.

A sentence containing items and activities from a fixed vocabulary was given to 267 different users, who then produced a video acting out the sentence. Its complexity and diversity make it a useful dataset for evaluating the ability of video captioning models to generate captions for a wide range of activities.

From Table 15, several conclusions can be drawn. First, we can interpret that the most frequently used datasets belong to primary categories of the open world, social media, and movies. Second, other than a few datasets that incorporate many phrases or even paragraphs per video sample, the majority of datasets only allocate one caption per video. Third, this table provides evidence that undoubtedly the size of the dataset has an impact on how well the method of video captioning performs. Fourth, short video duration can be easily handled by the model and its description is much easier. This could be the possible reason to select the MSVD and MSR-VTT datasets for benchmarking. So, similar trends are observed in the selection of both datasets.

The effectiveness of captioning models is significantly influenced by the number of natural language sentences that are provided for each video clip in a dataset. These sentences play a crucial role in shaping the output of the models and therefore the number of available sentences has a significant impact on their performance. Table 15 displays the sentence-to-word and video clip-to-sentence ratios for the selected datasets. With 40 sentences per video clip, MSVD has the highest ratio of sentences to video clips. A video captioning model can understand more events due to the open domain dataset's variety of activity classes. Figure 8 demonstrates how the MSR-VTT dataset in addition to MSVD has also been extensively used. However, when comparing the two datasets, the MSR-VTT dataset was found to have a larger number of video clips, sentences, and vocabulary compared to the MSVD dataset. However, when it came to evaluating the performance of different models, the MSVD dataset performed better and was therefore used more frequently, as shown in Figure 8. Moreover, the MSVD dataset contains videos that showcase a single event per clip, with an average length of 10 seconds per video. However, because MSR-VTT clips are typically 20 seconds long, several events overlap, making it difficult for the spatiotemporal feature extractor to discern one event from another.

Traditional benchmark datasets often have a small number of movies and a low level of narration, and the visual and textual material typically have simple semantics. To effectively utilize deep learning in the task of video captioning, various large and well-labeled datasets have been created. This allows for the potential to produce captions that are on par with human-level descriptions. Another aspect of video captioning is the ability to describe a video's content in multiple languages, which is referred to as multilingual video captioning. Most of the large-scale video captioning datasets that are available right now only support the English language, hence English corpora are the only ones that can be employed to create video captioning models. For the world's enormous non-English speaking population, the study of multilingual video captioning is crucial which is another novel task introduced by the new video captioning datasets. YouCookII, TRECVID, DVS, TACoS-MultiLevel, and VATEX are some other datasets that were only infrequently used in the selected studies. All of these datasets were unable to work effectively as high as MSVD or MSR-VTT.

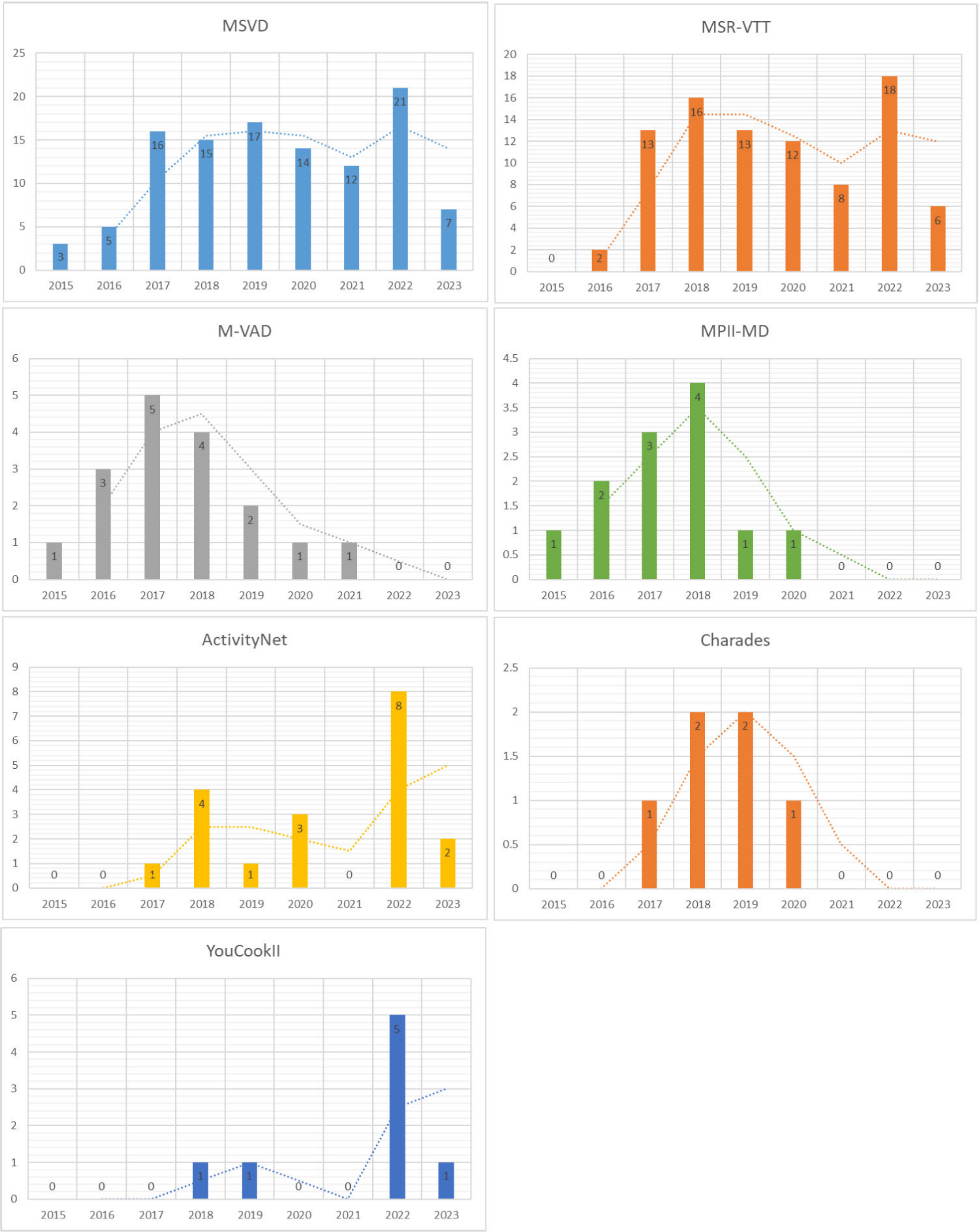

FIGURE 9. Year-wise distribution of datasets for selected primary studies.

The dataset domain is a key consideration in selecting it. Every other dataset, except ActivityNet and MSR-VTT, belongs to a specific domain. Very few activity classes exist in domain-specific datasets, which restricts a model's potential strength. For instance, the movie datasets M-VAD and MPII-MD, and the cooking dataset TACoS-Multi-Level.

Table 16 summarizes the characteristics of video captioning datasets that are most influential for the study of video captioning models including:

i. *Size:* A large size of the dataset is important to ensure that the video captioning models can be trained on a sufficient amount of data and generalize well to new videos.

The MSVD dataset is relatively small compared to more recent video captioning datasets, with only 1,970 video clips and captions. This makes it challenging to train video captioning models on the MSVD dataset, but it is still widely used due to its early use and its simplicity. MSR-VTT is a relatively large dataset, with 10,000 video clips and captions in 20 different languages, making it a popular choice for training and evaluating video captioning models. M-VAD and ActivityNet Captions are also large-scale video captioning datasets with this influential characteristic of size making them useful resources for training and evaluating video captioning models. MPII-MD and Charades are relatively small datasets for fine-tuning models pre-trained on larger datasets.

ii. *Diversity:* The dataset should contain a diverse range of videos and captions to represent different styles, languages, and subjects. This is important to ensure that the video captioning models are capable of generating captions for a wide range of videos and can generalize well to new videos.
The MSVD dataset contains a diverse range of videos, but the captions are relatively short and may not provide enough information to describe the full content of the videos. MSR-VTT and Charades are diverse datasets, with a wide range of video topics and captions in multiple languages, making them good benchmarks for evaluating the generalization ability of video captioning models. M-VAD contains a wide range of video topics and captions, making it a good benchmark for evaluating the generalization ability of video captioning models.

iii. *Quality:* The quality of the captions in the dataset is important because the captions serve as the ground truth for evaluating the performance of the video captioning models. Captions that are poorly written or do not accurately describe the content of the videos can negatively impact the evaluation of the video captioning models.
The quality of the captions in the MSVD dataset is generally good, but some captions may be inaccurate or incomplete. The quality of the captions in the MSR-VTT dataset is generally good, with accurate and descriptive captions that accurately describe the contents of the videos. The captions in MPII-MD are generally high-quality, making it a useful resource for evaluating the quality of the captions generated by video captioning models. The quality of the captions in the ActivityNet and M-VAD datasets is subjective and can depend on individual criteria and preferences. The captions in the Charades dataset are self-generated, meaning that they are produced by the participants themselves. As a result, the quality of the captions may vary and may not be equivalent to those produced by professional annotators. However, the self-generated nature of the captions in Charades also provides valuable information about how individuals describe their actions, which can be useful for certain research purposes.

iv. *Annotated information:* The dataset should contain additional annotated information, such as action labels, object labels, and scene labels, to allow for a more comprehensive evaluation of the video captioning models. M-VAD contains annotations for actions, objects, and scenes in the videos, making it a useful dataset for evaluating the ability of video captioning models to generate captions that describe not only what is happening in the video, but also how it is happening.

v. *Temporal information:* The dataset should contain information about the temporal structure of the videos and the captions, such as start and end times for actions, objects, and scenes, to allow for a more precise evaluation of the video captioning models.
The MSVD dataset does not contain any temporal information about the captions or the videos, making it difficult to evaluate the ability of the video captioning models to generate captions that accurately describe the timing of events in the videos. MSR-VTT, MPII-MD, and Charades provide temporal information about the captions and the videos to generate captions that accurately describe the timing of events in the videos.

vi. *Multi-modality:* The dataset should contain multiple modalities of information, such as audio, visual, and text, to allow for a more comprehensive evaluation of the video captioning models. M-VAD and Charades are multi-modal datasets, providing captions, actions, objects, and scene annotations for the videos, making them useful resources for evaluating the ability of video captioning models to generate captions that accurately describe the contents of the videos.

vii. *Evaluation protocol:* ActivityNet Captions provides a well-defined evaluation protocol, making it easier to compare the performance of different video captioning models on this dataset.

D. RESULTS OF RQ3: WHICH PERFORMANCE EVALUATION METRIC IS MOST APPROPRIATE FOR THE STUDIES OF VIDEO CAPTIONING MODELS RELATIVE TO PARAMETER CHARACTERISTICS?

To gauge how closely a model's video captioning resembles the human annotation, evaluation metrics are crucial in this task. Since there is no pre-defined correct answer or ground truth that can be used as a standard for measuring accuracy, evaluating video captioning is likewise difficult. Many different sentences, with differences in both syntactic structure and semantic content, can accurately describe a video clip. To address the 3rd RQ, the proportion of the performance evaluation metrics applied frequently in the selected studies is presented in Figure 10 and Figure 11 which could portray the performance of the method proposed. A total of four performance evaluation parameters were applied in the selected studies with significant frequency. The fact that the BLEU and METEOR are two of the most well-known performance evaluation measures, followed by the CIDEr and the ROUGE, is fairly remarkable to observe from Figure 11. Other than the above-mentioned metrics, some more metrics were applied rarely, e.g., SPICE, perplexity, precision, recall, and SVO Accuracy hardly implemented once or twice in the selected primary studies.

TABLE 16. Influential characteristics of video captioning datasets.

Dataset	Size	Videos diversity	Captions quality	Annotated information	Temporal information	Multi-modality	Evaluation protocol
MSVD	Small	✓	Good	×	×	×	×
MSR-VTT	Large	✓	Good	×	✓	×	×
M-VAD	Large	✓	-	✓	✓	✓	×
MPII-MD	Small	✓	High	×	✓	×	×
ActivityNet	Large	✓	-	×	✓	×	✓
Charades	Small	✓	-	×	✓	✓	×

In the task of video captioning, performance evaluation metrics are influenced by several characteristics. These characteristics can help to evaluate the performance of video captioning models and guide the development of better models. Relevance measures how closely the predicted caption matches the content of the video. A good video captioning model should generate captions that accurately describe the events and actions in the video. A good evaluation metric should be highly correlated with human judgments of the quality of machine-generated captions. The closer the evaluation metric is to human judgment, the more accurate it is in measuring the performance of the model. Fluency measures the naturalness and grammatical correctness of the predicted caption. A good video captioning model should generate captions that are fluent and readable. Coherence measures how well the predicted caption fits into a larger context and how well it follows a logical flow of events. A good video captioning model should generate captions that are coherent and have a logical structure. Diversity measures the range of different captions that can be generated for a given video. A good video captioning model should generate captions that are diverse and capture different aspects of the video content. This measures the consistency of the predicted captions across different videos. A good video captioning model should generate captions that are consistent and have a similar level of quality regardless of the video content. Table 15 lists them along with their frequency and original purpose of design. The interpretation of Figure 11 and Figure 10 shows that most used evaluation parameter applied so far is the METEOR and the least used is ROUGE.

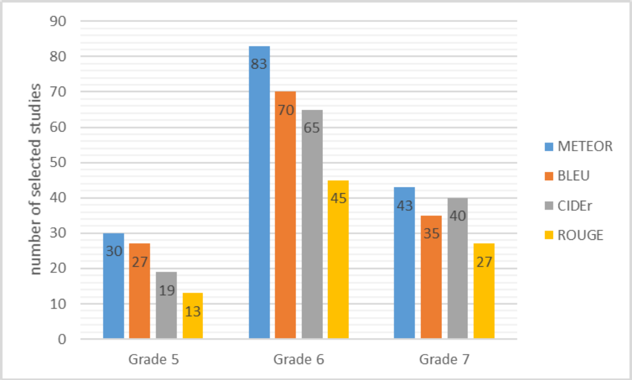

FIGURE 10. Grade-wise distribution of performance evaluation metrics.

In Table 17, the terms "high," "moderate," and "low" serve as qualitative indicators describing the observed levels of various performance evaluation parameters. When applied to the parameter of "Relevance," a designation of "high" implies a strong association or alignment with the original purpose of the evaluation, indicating a significant degree of relevance. Conversely, "moderate" suggests a moderate level of alignment. In the context of "Correlation with human judgments," "high" denotes a substantial correlation between the evaluation metric and human judgments, affirming the reliability of the metric. "Moderate" suggests a noticeable yet less pronounced correlation. For parameters such as "Fluency," "Coherence," and "Consistency," a characterization of "high" signifies a commendable level of the respective attribute, while "low" indicates a lesser degree of fluency, coherence, or consistency. These qualitative descriptors provide succinct insights into the observed performance characteristics across different evaluation parameters.

Its combination of word-level and sentence-level similarity measures makes it a comprehensive metric that considers multiple aspects of the generated captions, including accuracy, fluency, and coherence. METEOR also considers the word order and synonymy between the reference captions and the generated captions, which is important in video captioning, where captions must accurately reflect the content of the video and be grammatically correct. ROUGE is a widely used evaluation metric in the field of video captioning, as it measures the similarity between generated captions and reference captions. It calculates the overlap between n-grams of the generated and reference captions, including unigrams, bigrams, and trigrams, providing a comprehensive evaluation of the caption quality and coherence.

Here is the analysis of the mentioned characteristics of the selected evaluation metrics.

i. *Relevance:* In terms of relevance to video captioning, METEOR is considered to be a good choice because it measures both content-related and language-related aspects of the captions.

TABLE 17. List of performance evaluation parameters.

Evaluation parameter	METEOR	BLEU	CIDEr	ROUGE
Original Purpose	Machine translation	Machine translation	Image captioning	Document summarization
# of times applied in selected studies	156	132	125	86
Relevance	high	moderate	high	moderate
Correlation with human judgments	high	moderate	high	moderate
Fluency	high	low	high	low
Coherence	high	low	high	low
Diversity	high	low	-	-
Consistency	-	high	-	high

- indicates that the metric does not consider the specific characteristic.

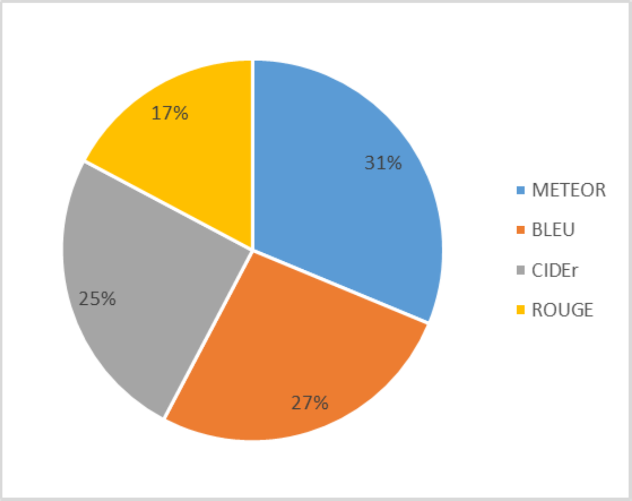

FIGURE 11. Proportion of performance evaluation metrics in selected studies

ROUGE is useful for benchmarking and comparing different video captioning models and helps to assess their ability to accurately describe the content of a video. In video captioning, BLEU can be used to assess the level of fluency and grammatical correctness of the generated captions. BLEU scores can range from 0 to 1, with higher scores indicating higher levels of similarity between the generated and reference captions. CIDEr calculates a weighted combination of n-gram precision and recalls, considering not only the overlapping n-grams but also the level of agreement between multiple human evaluators on the quality and relevance of the generated captions. CIDEr is considered to be a more effective metric than traditional n-gram-based metrics like BLEU and ROUGE, as it incorporates human evaluation into the scoring process. This makes CIDEr a valuable tool for evaluating the quality and relevance of video captions generated by machine learning models.

ii. *Correlation with human judgments:* The correlation of BLEU, CIDEr, ROUGE, and METEOR with human judgments of text generation and summarization quality is moderate. BLEU measures the n-gram overlap between an automatically generated text and a reference text created by a human, with higher BLEU scores indicating greater overlap. CIDEr measures the similarity between the automatically generated caption and a reference caption created by a human, with higher CIDEr scores indicating greater similarity. ROUGE measures the overlap between the automatically generated summary and a reference summary created by a human, and higher ROUGE scores indicate greater overlap. METEOR measures the harmonic mean of unigram precision and recall between the generated text and reference text, considering synonymy, stemming, and other factors.

While these metrics can be useful tools for evaluating text generation and summarization systems, it is important to remember that they do not consider all aspects of a good text from a human perspective, such as meaning, coherence, fluency, creativity, and readability. As a result, these metrics should not be used as the sole criterion for evaluating text generation or summarization quality. It is recommended to consider a combination of these metrics and human judgments to get a more complete picture of the performance of a text generation or summarization system.

iii. *Fluency:* BLEU is considered a good metric for measuring fluency, as it calculates the n-gram overlap between the generated captions and the reference captions. Higher BLEU scores indicate that the generated captions have a higher degree of overlap with the reference captions, which is often associated with fluency. METEOR also considers word alignments, synonymy, and paraphrase information in its similarity score calculation, so it can provide a better measure of the fluency of the generated captions than BLEU alone. However, it is not specifically designed to measure fluency and may not provide as accurate a measure of fluency as BLEU. CIDER and ROUGE, on the other hand, are not as well suited to measure fluency. CIDER measures the cosine similarity between the captions, which does not specifically capture fluency information. ROUGE measures the overlap between n-grams in the captions, which can indicate fluency to some degree, but does not provide a comprehensive measure of fluency. In conclusion, while all four metrics can provide some information about the fluency of the generated captions, BLEU is the most commonly used metric for evaluating fluency in video captioning.

iv. *Coherence:* The coherence of BLEU, METEOR, CIDER, and ROUGE for the task of video captioning evaluation refers to the degree to which the metrics accurately capture the semantical and logical consistency of the generated captions. BLEU is not designed to measure the coherence of the generated captions but rather focuses on n-gram overlap between the generated captions and reference captions. So, its coherence for the task of video captioning evaluation may be limited. METEOR incorporates word alignment and synonymy information into its similarity score calculation, which makes it a better metric for capturing semantic information and coherence of the generated captions compared to BLEU. CIDER measures the cosine similarity between the captions, which can be a good indicator of the semantic consistency and coherence of the generated captions. ROUGE measures the recall-oriented similarity between the generated captions and reference captions, which does not necessarily capture the semantic consistency and coherence of the generated captions. So, while all four metrics have different strengths and limitations, METEOR and CIDER tend to be better suited for capturing the coherence of the generated captions compared to BLEU and ROUGE.

v. *Diversity:* BLEU measures the n-gram overlap between the generated caption and reference captions, making it a less robust metric for measuring diversity. On the other hand, METEOR uses a combination of unigram precision, recall, and harmonic mean, which considers the overall meaning of the sentence, thus making it a better metric for measuring diversity in video captioning. CIDEr is a dense captioning metric that considers both word-overlap and semantic similarity, making it a robust metric for measuring diversity in video captioning. Lastly, ROUGE measures the recall and precision between the generated caption and reference captions, however, it does not consider the semantic similarity, making it a less robust metric for measuring diversity in video captioning.

vi. *Consistency:* The consistency of metrics such as BLEU, METEOR, CIDEr, and ROUGE for evaluating video captioning models can vary depending on the specific dataset used. Each of these metrics has its strengths and weaknesses, and the suitability of a particular metric for a given task can depend on the characteristics of the dataset in question. For example, BLEU is a widely used metric that measures the overlap between predicted captions and ground-truth captions, but it can be biased towards models that generate generic or common phrases, regardless of their relevance to the video content. BLEU is widely used in video captioning tasks, but it can be unreliable in cases where the machine-generated captions contain new and creative expressions that are not present in the ground truth captions. Whereas, Cider assigns a score based on the quality of the captions and not just on their consistency with the ground truth captions. In summary, the suitability of a particular metric will depend on the specific characteristics of the video captioning dataset, and it is important to use multiple metrics for a comprehensive understanding of model performance.

It is important to note that the metrics mentioned above are not perfect and each has its strengths and weaknesses, and choosing the appropriate metric for the video captioning task may depend on the specific requirements of the task and the dataset being evaluated. Additionally, the results of these metrics should not be used in isolation, and a combination of metrics may provide a more comprehensive evaluation of the generated captions.

Various machine translation tasks are analyzed and tested using a variety of performance evaluation measures. However, the research claims that a significant limitation is that there isn't a performance evaluation metric dedicatedly designed for measuring video captioning, hence this task has instead been accomplished using several machine translation and image captioning tasks.

III. CONCLUSION

This study presents a systematic review of video captioning studies representing almost all the required details. The paper has explored all the possible methods of machine learning and deep learning used for visual feature extraction and language processing. A total of 162 primary studies were selected to answer the research question based on the defined selection criteria. Statistical analysis of all these state-of-the-art methods has been presented. The hybrid technique was mostly employed for both feature extraction and caption generation sub-tasks. 2D/3D CNN methods were the most frequently used; 57, and 44 times respectively for the feature extraction. LSTM was 96 times employed for caption generation. MSVD and MSR-VTT are the most frequently adopted datasets applied in 96, and 73 studies respectively. Subsequently, we have identified all evaluation metrics applied to verify the accuracy of generated captions of videos. METEOR being used 156 times and BLEU used 132 times were identified as the most frequently applied performance evaluation parameters, respectively. We have found that EMScore is the only metric recently designed to evaluate the accuracy of generated captions for videos. Rather, the metrics dedicated to image captioning and machine translation domain are being used for video captioning. One limitation observed in most of the primary studies is that the comparison of results is conducted with older studies, whereas newer studies consistently demonstrate significantly higher results. Nonetheless, a few research limitations associated with the conducted study are related to the search duration used in all the databases. the methodology section describes that the studies span from database incision to October 07, 2023. The studies published after this date were excluded. Another limiting factor is the database limitation. The record retrieval is strictly limited to the four scientific databases only. Thus, based on the results of this systematic review, the above trends are considered obvious. Adding some more databases would benefit enhancing the search process. We also believe that it is also important to compare the study results with state-of-the-art studies. Lastly, the accuracy level of methods would be truly judged by the development of original video captioning metrics. In conclusion, this combinatorial analysis of deep learning and machine learning video captioning studies provides valuable insights into the current state of the art in this field and helps in determining the best approaches, datasets, and performance evaluation metrics for video captioning analyzing their key characteristics.

IV. FUTURE AGENDA

In this study, a thorough examination of the existing literature and research in the field of video captioning was conducted. Significant achievements have been made with this recent joint venture of visual recognition and computational linguistics. The results identify that further research is needed by highlighting the following outcomes and promising directions:

- One of the key outcomes of this systematic review is the recognition of the need for more comprehensive databases and primary studies to be conducted. This will enable researchers to explore a wider range of methods and data and expand the body of knowledge in the field.
- Additionally, this systematic review provides a solid foundation for a potential meta-analysis, which would bring together the findings of selected studies using statistical techniques.

Overall, this systematic review provides a comprehensive overview of the video captioning field and serves as a valuable resource for researchers, practitioners, and students alike.

APPENDIX

REVIEW PROTOCOL: PROTOCOL FOR A SYSTEMATIC LITERATURE REVIEW ON VIDEO CAPTIONING STUDIES

A. BACKGROUND

In 1991, Koller et al. [1] came up with an innovative approach for describing the movements of vehicles in real-world traffic situations using verbs from natural language. Koller's study was followed up by Brand et al. in 1997, who created a storyboard from instructional videos and summarized a sequence of events into semantic tag descriptions. Following these works, Kojima et al. [2] produced a phrase based on predefined templates in 2002 after first acquiring visual conceptions in a video clip with handcrafted features. A new age of video description started in 2011 [3] when models started to get more complex and adaptable. The objective of this study is to conduct a comprehensive review of the existing research on video captioning, with a focus on the various methods used for generating both single-sentence and dense captions for videos. The aim is to provide a comprehensive overview of the state-of-the-art in video captioning and identify any gaps in the existing literature that may warrant further investigation.

TABLE 18. Sources to be searched.

Source	Responsible
Scopus	Author 1
Web of Science	Author 2
IEEE Xplore	Author 3, Author 4
ACM	Author 5

B. RESEARCH QUESTIONS

The following questions will be the focus of the investigation:

- What evidence indicates that deep learning methods are more commonly utilized than machine learning methods in video captioning?
- Which video captioning datasets relative to different parameter characteristics are used preferably for the benchmarking?
- Which performance evaluation metric is most appropriate for the studies of video captioning models relative to parameter characteristics?

C. SEARCH PROCESS

This study conducted a manual search of specific conference proceedings and journal papers from the inception of relevant databases. The selected databases are listed in a table.

D. INCLUSION CRITERIA

Studies exhibiting the following characteristics, published between database commencement and September 30th, 2022, will be included:

- Primary studies with a defined research objective of video captioning
- Downloadable studies

E. EXCLUSION CRITERIA

This study will not consider certain types of papers such as:

- Inappropriate studies due to no defined research methodology
- Review studies
- Studies written in a non-English language

F. STUDY SELECTION

The process of selecting studies will be carried out in two stages: first, the titles and abstracts will be examined against the inclusion criteria to find studies that might be relevant, and then the full papers of those studies that passed the initial screening will be evaluated. The titles and abstracts will be evaluated by authors 1 and 2, and their evaluation will be reviewed by authors 3 and 4. The full papers will then be studied by authors 3 and 4, and authors 1 and 2 will assess their work. Conflicts over study eligibility shall be settled amicably. Author5 will review the rejected studies and tabulate the results as follows:

- Annual selection of studies by source
- The total number of articles chosen each year

G. QUALITY ASSESSMENT

Primary studies will be evaluated under the criteria which are based on the following seven questions:

- Does the study give a clear description of its goal?
- Is there a publicly available dataset used in the study?
- Does the study mention where the data came from?
- Is the study using appropriate performance evaluation metrics for the analysis of the results?
- Were the outcomes contrasted with findings from earlier research?
- Was a comparison of the results to the most recent studies done?
- Does the conclusion have any relevance to the study's goal?

The questions are scored as follows:

- *Question 1:* Y (yes), the research objective is explicitly defined in the paper; P (Partly), the research objective is implicit; N (no), the research objective is not defined and cannot be readily inferred.
- *Question 2:* Y (yes), the standard datasets have been adopted; P, data augmentation or some other way has been applied on standard datasets or a mix of some custom-built dataset, and standard datasets have been used; N, private or custom-built datasets have been used.
- *Question 3:* Y (yes), any type of primary, secondary, or tertiary source is mentioned for the dataset; P, the dataset references implicitly any kind of primary, secondary, or tertiary source; N no reference of any kind of primary, secondary, or tertiary source mentioned for the dataset.
- *Question 4:* Y (yes), the study evaluates its performance using all the METEOR, BLEU, CIDEr, or ROUGE variants; P, the study assesses its effectiveness using some variants of METEOR, BLEU, CIDEr, or ROUGE; N no variant of METEOR, BLEU, CIDEr, or ROUGE employed, rather used some other metrics, or even didn't use any metric.
- *Question 5:* Y (yes), the study comparing the outcomes to other benchmark studies in proper tabulation; P, the study comparing the outcomes to unfamiliar studies or some generic models; N the study not comparing the outcomes to other studies.
- *Question 6:* Y (yes), the most recent studies were used to compare the results, of the same year, or one year older; P, two-year older studies were used to compare the results; N older than two-year studies were used to compare the results.
- *Question 7:* Y (yes), the conclusion is fully related to and mapped with the research objectives; P, the conclusion is somehow related to and mapped with the research objectives; N the conclusion doesn't discuss the research objectives.

The scoring procedure is Y=1, P=0.5, and N or Unknown =0.

One researcher will extract the data, and another will check it.

H. DATA COLLECTION

The data extracted from each paper will be:

- document type (i.e. the conference paper or journal article)
- year of paper publication
- classification of proposed methodology (i.e., features extraction model, and caption generation model)
- datasets
- performance metrics
- number of citations
- venue title
- quality score for the study

The task of data collection will be carried out by a single researcher, who will be responsible for extracting all the relevant information from the sources. This collected data will then be reviewed and verified by another researcher to ensure its accuracy and completeness. This double-checking process will ensure the validity and reliability of the data.

I. DATA ANALYSIS

The data will be tabulated (ordered in reverse chronology) to show the quality grades of each study. There will be a count of the studies in each main category. To respond to the research objectives and spot any noteworthy patterns or constraints in the study of video captioning, the gathered data will be examined and plotted.

J. DISSEMINATION

Researchers who are interested in video captioning as well as those who work in computer vision and natural language processing might find the study's findings interesting. We, therefore, want to record, report, and publish the complete findings of the study.

K. REFERENCES

1. D. Koller, N. Heinze, and H. H. Nagel, "Algorithmic characterization of vehicle trajectories from image sequences by motion verbs," in Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1991.

2. A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," International Journal of Computer Vision, vol. 50, no. 2, pp. 171–184, 2002.

3. M. U. G. Khan, L. Zhang, and Y. Gotoh, ''Human-focused video description,'' in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops, IEEE, 2011.

AUTHOR CONTRIBUTIONS

Conceptualization and methodology; formal analysis, Wael M. S. Yafooz and Arafat Al-Dhaqm; writing and editing, Tanzila Kehkashan; visualization, Wael M. S. Yafooz; review and supervision, Abdullah Alsaeedi and Nor Azman Ismail. All authors have read and agreed to the published version of the manuscript.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

THE CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] 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, Nov. 2020.

[2] A. Puscasiu, A. Fanca, D.-I. Gota, and H. Valean, ''Automated image captioning,'' in *Proc. IEEE Int. Conf. Autom., Quality Test., Robot. (AQTR)*, May 2020, pp. 1–6.

[3] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, ''Deep reinforcement learning-based image captioning with embedding reward,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1151–1159.

[4] V. Jain, F. Al-Turjman, G. Chaudhary, D. Nayar, V. Gupta, and A. Kumar, ''Video captioning: A review of theory, techniques and practices,'' *Multimedia Tools Appl.*, vol. 81, no. 25, pp. 35619–35653, Oct. 2022.

[5] A. Kojima, T. Tamura, and K. Fukunaga, ''Natural language description of human activities from video images based on concept hierarchy of actions,'' *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, 2002.

[6] 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,'' 2012, *arXiv:1204.2742*.

[7] 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.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens, ''Moses: Open source toolkit for statistical machine translation,'' in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics Companion*, 2017, pp. 177–180.

[9] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2634–2641.

[10] 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. German Conf. Pattern Recognit.*, 2014, pp. 184–195.

[11] J. Corso, ''GBS: Guidance by semantics-using high-level visual inference to improve vision-based mobile robot localization,'' State Univ. New York (SUNY) Buffalo, Buffalo, NY, USA, Tech. Rep., 2015.

[12] 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.*, Oct. 2015, pp. 4534–4542.

[13] S. Islam, A. Dash, A. Seum, A. H. Raj, T. Hossain, and F. M. Shah, ''Exploring video captioning techniques: A comprehensive survey on deep learning methods,'' *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–28, Apr. 2021.

[14] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, and S. E. Brennan, ''The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,'' *Int. J. Surg.*, vol. 88, Apr. 2021, Art. no. 105906.

[15] B. Kitchenham, ''Procedures for performing systematic reviews,'' *Keele, UK, Keele Univ.*, vol. 33, no. 2004, pp. 1–26, 2004.

[16] L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, and L. A. Stewart, ''Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation,'' *Bmj*, vol. 349, Jan. 2015, Art. no. g7647.

[17] A. Cooke, D. Smith, and A. Booth, ''Beyond PICO: The SPIDER tool for qualitative evidence synthesis,'' *Qual. Health Res.*, vol. 22, no. 10, pp. 1435–1443, 2012.

[18] E. Stefana, F. Marciano, D. Rossi, P. Cocca, and G. Tomasoni, ''Wearable devices for ergonomics: A systematic literature review,'' *Sensors*, vol. 21, no. 3, p. 777, Jan. 2021.

[19] L. M. Kmet, L. S. Cook, and R. C. Lee, ''Standard quality assessment criteria for evaluating primary research papers from a variety of fields,'' Health Technol. Assessment Unit, Alberta Heritage Found. Medical Res., Edmonton, AB, Canada, Tech. Rep., 2004.

[20] U. Sirisha and B. S. Chandana, ''GITAAR-GIT based abnormal activity recognition on UCF crime dataset,'' in *Proc. 5th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Jan. 2023, pp. 1585–1590.

[21] J. Huang, H. Yan, L. Liu, and Y. Liu, ''Video description method with fusion of instance-aware temporal features,'' in *Proc. 3rd Int. Conf. Image Process. Intell. Control (IPIC)*, Aug. 2023, pp. 30–35.

[22] D. Rothenpieler and S. Amiriparian, ''METEOR guided divergence for video captioning,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–7.

[23] Y. Zheng, H. Jing, Q. Xie, Y. Zhang, R. Feng, T. Zhang, and S. Gao, ''Video captioning via relation-aware graph learning,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[24] M. Nabati and A. Behrad, ''Multi-sentence video captioning using spatial saliency of video frames and content-oriented beam search algorithm,'' *Exp. Syst. Appl.*, vol. 228, Oct. 2023, Art. no. 120454.

[25] S. Jing, H. Zhang, P. Zeng, L. Gao, J. Song, and H. T. Shen, ''Memory-based augmentation network for video captioning,'' *IEEE Trans. Multimedia*, vol. 26, pp. 2367–2379, 2024.

[26] X. Huang, K.-H. Chan, W. Wu, H. Sheng, and W. Ke, ''Fusion of multi-modal features to enhance dense video caption,'' *Sensors*, vol. 23, no. 12, p. 5565, Jun. 2023.

[27] M. S. Zaoad, M. M. R. Mannan, A. B. Mandol, M. Rahman, M. A. Islam, and M. M. Rahman, ''An attention-based hybrid deep learning approach for Bengali video captioning,'' *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 1, pp. 257–269, Jan. 2023.

[28] T.-Z. Niu, S.-S. Dong, Z.-D. Chen, X. Luo, S. Guo, Z. Huang, and X.-S. Xu, ''Semantic enhanced video captioning with multi-feature fusion,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 6, pp. 1–21, Nov. 2023.

[29] H. Wang, P. Tang, Q. Li, and M. Cheng, ''Emotion expression with fact transfer for video description,'' *IEEE Trans. Multimedia*, vol. 24, pp. 715–727, 2022.

[30] D. T. Phuc, T. Q. Trieu, N. V. Tinh, and D. S. Hieu, ''Video captioning in Vietnamese using deep learning,'' *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 12, no. 3, p. 3092, Jun. 2022.

[31] N. Aafaq, A. Mian, W. Liu, N. Akhtar, and M. Shah, ''Cross-domain modality fusion for dense video captioning,'' *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 763–777, Oct. 2022.

[32] S. Varma and J. D. Peter, ''Deep learning-based video captioning technique using transformer,'' in *Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2022, pp. 847–850.

[33] J. Madake, S. Bhatlawande, S. Purandare, S. Shilaskar, and Y. Nikhare, ''Dense video captioning using BiLSTM encoder,'' in *Proc. 3rd Int. Conf. Emerg. Technol. (INCET)*, May 2022, pp. 1–6.

[34] H. Xiao and J. Shi, ''Diverse video captioning through latent variable expansion,'' *Pattern Recognit. Lett.*, vol. 160, pp. 19–25, Aug. 2022.

[35] A. Singh, T. D. Singh, and S. Bandyopadhyay, ''V2T: Video to text framework using a novel automatic shot boundary detection algorithm,'' *Multimedia Tools Appl.*, vol. 81, no. 13, pp. 17989–18009, May 2022.

[36] D. Naik and C. D. Jaidhar, "A novel multi-layer attention framework for visual description prediction using bidirectional LSTM," *J. Big Data*, vol. 9, no. 1, p. 104, Nov. 2022.

[37] J. Prudviraj, M. I. Reddy, C. Vishnu, and C. K. Mohan, "AAP-MIT: Attentive atrous pyramid network and memory incorporated transformer for multisentence video description," *IEEE Trans. Image Process.*, vol. 31, pp. 5559–5569, 2022.

[38] J. Jacob and V. P. Devassia, "Dense captioning of videos using feature context integrated deep LSTM with local attention," in *Proc. 6th Int. Conf. I-SMAC*, Nov. 2022, pp. 810–818.

[39] Y. Yuan, L. Ma, and W. Zhu, "Syntax customized video captioning by imitating exemplar sentences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10209–10221, Dec. 2022.

[40] C. Chatzikonstantinou, G. G. Valasidis, K. Stavridis, G. Malogiannis, A. Axenopoulos, and P. Daras, "UCF-CAP, video captioning in the wild," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 1386–1390.

[41] Y. Tu, L. Li, L. Su, S. Gao, C. Yan, Z.-J. Zha, Z. Yu, and Q. Huang, "I2Transformer: Intra- and inter-relation embedding transformer for TV show captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 3565–3577, 2022.

[42] E. Mavroudi and R. Vidal, "Weakly-supervised generation and grounding of visual descriptions with conditional generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15523–15533.

[43] L. Li, Y. Zhang, S. Tang, L. Xie, X. Li, and Q. Tian, "Adaptive spatial location with balanced loss for video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 17–30, Jan. 2022.

[44] J. Vaidya, A. Subramaniam, and A. Mittal, "Co-segmentation aided two-stream architecture for video captioning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2442–2452.

[45] C.-H. Lu and G.-Y. Fan, "Environment-aware dense video captioning for IoT-enabled edge cameras," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4554–4564, Mar. 2022.

[46] L. Li, X. Gao, J. Deng, Y. Tu, Z.-J. Zha, and Q. Huang, "Long short-term relation transformer with global gating for video captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 2726–2738, 2022.

[47] D. Naik and C. D. Jaidhar, "Semantic context driven language descriptions of videos using deep neural network," *J. Big Data*, vol. 9, no. 1, pp. 1–22, Dec. 2022.

[48] Y. Zheng, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Stacked multimodal attention network for context-aware video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 31–42, Jan. 2022.

[49] H. Im and Y.-S. Choi, "UAT: Universal attention transformer for video captioning," *Sensors*, vol. 22, no. 13, p. 4817, Jun. 2022.

[50] T. Deb, A. Sadmanee, K. K. Bhaumik, A. A. Ali, M. A. Amin, and A. K. M. M. Rahman, "Variational stacked local attention networks for diverse video captioning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2493–2502.

[51] P. Tang, Y. Tan, and W. Luo, "Visual and language semantic hybrid enhancement and complementary for video description," *Neural Comput. Appl.*, vol. 34, no. 8, pp. 5959–5977, Apr. 2022.

[52] R. S. Bhooshan and K. Suresh, "A multimodal framework for video caption generation," *IEEE Access*, vol. 10, pp. 92166–92176, 2022.

[53] B. Wu, G. Niu, J. Yu, X. Xiao, J. Zhang, and H. Wu, "Towards knowledge-aware video captioning via transitive visual relationship detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6753–6765, Oct. 2022.

[54] X. Hua, X. Wang, T. Rui, F. Shao, and D. Wang, "Adversarial reinforcement learning with object-scene relational graph for video captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 2004–2016, 2022.

[55] S. Li, B. Yang, and Y. Zou, "Utilizing text-based augmentation to enhance video captioning," in *Proc. 5th Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2022, pp. 287–293.

[56] H. Zhao, Z. Chen, L. Guo, and Z. Han, "Video captioning based on vision transformer and reinforcement learning," *PeerJ Comput. Sci.*, vol. 8, p. e916, Mar. 2022.

[57] S. Li, Z.-F. Zhang, Y. Ji, Y. Li, and C.-P. Liu, "Spatio-temporal graph-based semantic compositional network for video captioning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[58] W. Choi, J. Chen, and J. Yoon, "Parallel pathway dense video captioning with deformable transformer," *IEEE Access*, vol. 10, pp. 129899–129910, 2022.

[59] P. Song, D. Guo, J. Cheng, and M. Wang, "Contextual attention network for emotional video captioning," *IEEE Trans. Multimedia*, vol. 25, pp. 1858–1867, 2022.

[60] N. Yadav and D. Naik, "Loss optimised video captioning using deep-LSTM, attention mechanism and weighted loss metrices," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol.*, 2021, pp. 1–7.

[61] R. Radarapu, A. S. S. Gopal, and N. H. Madhusudhan, "Video summarization and captioning using dynamic mode decomposition for surveillance," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 1927–1936, Oct. 2021.

[62] W. Ji and R. Wang, "A multi-instance multi-label dual learning approach for video captioning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 2s, pp. 1–18, Jun. 2021.

[63] J.-C. Lin and C.-Y. Zhang, "A new memory based on sequence to sequence model for video captioning," in *Proc. Int. Conf. Secur., Pattern Anal., Cybernetics(SPAC)*, Jun. 2021, pp. 470–476.

[64] H. Maru, T. Chandana, and D. Naik, "Comparitive study of GRU and LSTM cells based video captioning models," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2021, pp. 1–5.

[65] Y. Huang, L. Shih, C. Tsai, and G. Shen, "Describing video scenarios using deep learning techniques," *Int. J. Intell. Syst.*, vol. 36, no. 6, pp. 2465–2490, Jun. 2021.

[66] S. Liu, Z. Ren, and J. Yuan, "SibNet: Sibling convolutional encoder for video captioning," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1425–1434.

[67] W. Yan, L. Qi, Y. Tie, and C. Jin, "Video captioning via two-stage attention model and generative adversarial network," in *Proc. 8th Int. Conf. Comput. Science/Intell. Appl. Informat. (CSII)*, Sep. 2021, pp. 6–11.

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

[69] A. Wu, Y. Han, Z. Zhao, and Y. Yang, "Hierarchical memory decoder for visual narrating," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2438–2449, Jun. 2021.

[70] W. Tian and Y. Hu, "Label importance ranking with entropy variation complex networks for structured video captioning," *Traitement du Signal*, vol. 38, no. 4, pp. 937–946, Aug. 2021.

[71] H. Kim and S. Lee, "A video captioning method based on multi-representation switching for sustainable computing," *Sustainability*, vol. 13, no. 4, p. 2250, Feb. 2021.

[72] A. H. Raj, A. Seum, A. Dash, S. Islam, and F. M. Shah, "Deep learning based video captioning in Bengali," in *Proc. 26th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2021, pp. 1–6.

[73] W. Xu, J. Yu, Z. Miao, L. Wan, Y. Tian, and Q. Ji, "Deep reinforcement polishing network for video captioning," *IEEE Trans. Multimedia*, vol. 23, pp. 1772–1784, 2021.

[74] Y. Jin, J. Kwak, Y. Lee, J. Yun, and H. Ko, "KU-ISPL TRECVID 2018 VTT model," Health Technol. Assessment Unit, Alberta Heritage Found. Medical Res., Edmonton, AB, Canada, Tech. Rep., 2018.

[75] F. Zhu, J. Hwang, Z. Ma, G. Chen, and J. Guo, "Understanding objects in video: Object-oriented video captioning via structured trajectory and adversarial learning," *IEEE Access*, vol. 8, pp. 169146–169159, 2020.

[76] K. Ning, M. Cai, D. Xie, and F. Wu, "An attentive sequence to sequence translator for localizing video clips by natural language," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2434–2443, 2020.

[77] X. Hao, F. Zhou, and X. Li, "Scene-edge GRU for video caption," in *Proc. IEEE 4th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, vol. 1, Jun. 2020, pp. 1290–1295.

[78] L. Gao, X. Wang, J. Song, and Y. Liu, "Fused GRU with semantic-temporal attention for video captioning," *Neurocomputing*, vol. 395, pp. 222–228, Jun. 2020.

[79] X. Li, J. Ye, C. Xu, S. Yun, L. Zhang, X. Wang, R. Qian, and J. Dong, "Renmin University of China and Zhejiang Gongshang University at TRECVID 2019: Learn to search and describe videos," Health Technol. Assessment Unit, Alberta Heritage Found. Medical Res., Edmonton, AB, Canada, Tech. Rep., 2019.

[80] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "STAT: Spatial–temporal attention mechanism for video captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Jan. 2020.

[81] N. Kim, S. J. Ha, and J.-W. Kang, "Temporal attention feature encoding for video captioning," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 1279–1282.

[82] S. Sah, T. Nguyen, and R. Ptucha, ''Understanding temporal structure for video captioning,'' *Pattern Anal. Appl.*, vol. 23, no. 1, pp. 147–159, Feb. 2020.

[83] X. Long, C. Gan, and G. de Melo, ''Video captioning with multi-faceted attention,'' *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 173–184, Dec. 2018.

[84] M. Wang, ''Video description with GAN,'' in *Proc. IEEE 3rd Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, Aug. 2020, pp. 10–13.

[85] J. Kim, I. Choi, and M. Lee, ''Context aware video caption generation with consecutive differentiable neural computer,'' *Electronics*, vol. 9, no. 7, p. 1162, Jul. 2020.

[86] A. Wu, Y. Han, Y. Yang, Q. Hu, and F. Wu, ''Convolutional reconstruction-to-sequence for video captioning,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4299–4308, Nov. 2020.

[87] Z. Zhang, D. Xu, W. Ouyang, and C. Tan, ''Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3130–3139, Sep. 2020.

[88] A. Cherian, J. Wang, C. Hori, and T. K. Marks, ''Spatio-temporal ranked-attention networks for video captioning,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1606–1615.

[89] M. Qi, Y. Wang, A. Li, and J. Luo, ''Sports video captioning via attentive motion representation and group relationship modeling,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2617–2633, Aug. 2020.

[90] P. Tang, Y. Tan, J. Li, and B. Tan, ''Translating video into language by enhancing visual and language representations,'' *J. Vis. Commun. Image Represent.*, vol. 72, Oct. 2020, Art. no. 102875.

[91] Z. Liu, T. Chen, E. Ding, Y. Liu, and W. Yu, ''Attention-based convolutional LSTM for describing video,'' *IEEE Access*, vol. 8, pp. 133713–133724, 2020.

[92] J. Xu, H. Wei, L. Li, Q. Fu, and J. Guo, ''Video description model based on temporal–spatial and channel multi-attention mechanisms,'' *Appl. Sci.*, vol. 10, no. 12, p. 4312, Jun. 2020.

[93] N. Guo, H. Liu, and L. Jiang, ''Attention-based visual-audio fusion for video caption generation,'' in *Proc. IEEE 4th Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Jul. 2019, pp. 839–844.

[94] S. N. Aakur, F. D. de Souza, and S. Sarkar, ''Generating open world descriptions of video using common sense knowledge in a pattern theory framework,'' *Quart. Appl. Math.*, vol. 77, no. 2, pp. 323–356, Jan. 2019.

[95] S. Yang, W. Zhang, W. Lu, H. Wang, and Y. Li, ''Learning actions from human demonstration video for robotic manipulation,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 1805–1811.

[96] E. Ding, Z. Liu, Y. Liu, D. Xu, S. Feng, and X. Liu, ''Unsafe action recognition of miners based on video description,'' in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2019, pp. 1–4.

[97] Y. Wang, J. Liu, and X. Wang, ''Video description with integrated visual and textual information,'' *China Commun.*, vol. 16, no. 1, pp. 119–128, Jan. 2019.

[98] X. Li, Z. Zhou, L. Chen, and L. Gao, ''Residual attention-based LSTM for video captioning,'' *World Wide Web*, vol. 22, no. 2, pp. 621–636, Mar. 2019.

[99] Y. Zhu and S. Jiang, ''Attention-based densely connected LSTM for video captioning,'' in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 802–810.

[100] B. Zhao, X. Li, and X. Lu, ''CAM-RNN: Co-attention model based RNN for video captioning,'' *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, Nov. 2019.

[101] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, ''From deterministic to generative: Multimodal stochastic RNNs for video captioning,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.

[102] N. Laokulrat, N. Okazaki, and H. Nakayama, ''Incorporating semantic attention in video description generation,'' in *Proc. 11th Int. Conf. Language Resour. Eval.*, 2018, pp. 1–7.

[103] K. Gkountakos, A. Dimou, G. Th. Papadopoulos, and P. Daras, ''Incorporating textual similarity in video captioning schemes,'' in *Proc. IEEE Int. Conf. Eng., Technol. Innov. (ICE/ITMC)*, Jun. 2019, pp. 1–6.

[104] T. Fujii, Y. Sei, Y. Tahara, R. Orihara, and A. Ohsuga, '''Never fry carrots without cutting'. Cooking recipe generation from videos using deep learning considering previous process,'' in *Proc. IEEE Int. Conf. Big Data, Cloud Comput., Data Sci. Eng. (BCD)*, May 2019, pp. 124–129.

[105] J. Ye, L. Dong, W. Dong, N. Feng, and N. Zhang, ''Policy multi-region integration for video description,'' in *Proc. ACM Turing Celebration Conf.*, May 2019, pp. 1–5.

[106] X. Du, J. Yuan, L. Hu, and Y. Dai, ''Description generation of open-domain videos incorporating multimodal features and bidirectional encoder,'' *Vis. Comput.*, vol. 35, no. 12, pp. 1703–1712, Dec. 2019.

[107] S. Olivastri, G. Singh, and F. Cuzzolin, ''End-to-end video captioning,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Jun. 2019, pp. 8739–8748.

[108] 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.

[109] T. Jin, Y. Li, and Z. Zhang, ''Recurrent convolutional video captioning with global and local attention,'' *Neurocomputing*, vol. 370, pp. 118–127, Dec. 2019.

[110] Y. Xu, J. Yang, and K. Mao, ''Semantic-filtered soft-split-aware video captioning with audio-augmented feature,'' *Neurocomputing*, vol. 357, pp. 24–35, Sep. 2019.

[111] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, ''Memory-attended recurrent network for video captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8339–8348.

[112] J. Zhang and Y. Peng, ''Object-aware aggregation with bidirectional temporal graph for video captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8319–8328.

[113] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12479–12488.

[114] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, ''Streamlined dense video captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6581–6590.

[115] C. Hori, T. Hori, T. K. Marks, and J. R. Hershey, ''Early and late integration of audio features for automatic video description,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 430–436.

[116] F. Bolelli, L. Baraldi, F. Pollastri, and C. Grana, ''A hierarchical quasi-recurrent approach to video captioning,'' in *Proc. IEEE Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2018, pp. 162–167.

[117] X. Wu, G. Li, Q. Cao, Q. Ji, and L. Lin, ''Interpretable video captioning via trajectory structured localization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6829–6837.

[118] J. Yuan, C. Tian, X. Zhang, Y. Ding, and W. Wei, ''Video captioning with semantic guiding,'' in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–5.

[119] G. Wang, Z. Qin, K. Xu, K. Huang, and S. Ye, ''Bridge video and text with cascade syntactic structure,'' in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 3576–3585.

[120] S. Li, M. Tang, and J. Zhang, ''Deep hierarchical attention network for video description,'' *J. Electron. Imag.*, vol. 27, no. 2, 2018, Art. no. 023027.

[121] M. Bolaños, Á. Peris, F. Casacuberta, S. Soler, and P. Radeva, ''Egocentric video description based on temporally-linked sequences,'' *J. Vis. Commun. Image Represent.*, vol. 50, pp. 205–216, Jan. 2018.

[122] R. Shetty, H. R. Tavakoli, and J. Laaksonen, ''Image and video captioning with augmented neural architectures,'' *IEEE MultimediaMag.*, vol. 25, no. 2, pp. 34–46, Apr. 2018.

[123] S. Oura, T. Matsukawa, and E. Suzuki, ''Multimodal deep neural network with image sequence features for video captioning,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

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

[125] P.-H. Chang and A.-H. Tan, ''Learning generalized video memory for automatic video captioning,'' in *Proc. Int. Conf. Multi-Disciplinary Trends Artif. Intell.*, 2018, pp. 187–201.

[126] Y. Chen, W. Zhang, S. Wang, L. Li, and Q. Huang, ''Saliency-based spatiotemporal attention for video captioning,'' in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–8.

[127] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H. T. Shen, and Y. Ji, ''Video captioning by adversarial LSTM,'' *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.

[128] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4213–4222.

[129] A.-A. Liu, Y. Qiu, Y. Wong, Y.-T. Su, and M. Kankanhalli, "A fine-grained spatial–temporal attention model for video captioning," *IEEE Access*, vol. 6, pp. 68463–68471, 2018.

[130] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7190–7198.

[131] N. Xu, A.-A. Liu, Y. Wong, Y. Zhang, W. Nie, Y. Su, and M. Kankanhalli, "Dual-stream recurrent neural network for video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2482–2493, Aug. 2019.

[132] S. Cascianelli, G. Costante, T. A. Ciarfuglia, P. Valigi, and M. L. Fravolini, "Full-GRU natural language video description for service robotics applications," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 841–848, Apr. 2018.

[133] 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.

[134] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 358–373.

[135] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "M^3: Multimodal memory modelling for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7512–7520.

[136] Y. Xiong, B. Dai, and D. Lin, "Move forward and tell: A progressive generator of video descriptions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 468–483.

[137] C. Hori, T. Hori, G. Wichern, J. Wang, T.-Y. Lee, A. Cherian, and T. K. Marks, "Multimodal attention for fusion of audio and spatiotemporal features for video description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 2528–2531.

[138] S. Lee and I. Kim, "Multimodal feature learning for video captioning," *Math. Problems Eng.*, vol. 2018, pp. 1–8, Feb. 2018.

[139] 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.

[140] Y. Xu, Y. Han, R. Hong, and Q. Tian, "Sequential video VLAD: Training the aggregation locally and temporally," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4933–4944, Oct. 2018.

[141] S. Lee and I. Kim, "Video captioning with visual and semantic features," *J. Inf. Process. Syst.*, vol. 14, no. 6, pp. 1318–1330, 2018.

[142] S. Duggal, S. Manik, and M. Ghai, "Amalgamation of video description and multiple object localization using single deep learning model," in *Proc. 9th Int. Conf. Signal Process. Syst.*, Nov. 2017, pp. 109–115.

[143] H. Afli, F. Hu, J. Du, D. Cosgrove, K. McGuinness, N. E. O'Connor, E. A. Sánchez, J. Zhou, and A. F. Smeaton, "Dublin City University participation in the VTT track at TRECVid 2017," Health Technol. Assessment Unit, Alberta Heritage Found. Medical Res., Edmonton, AB, Canada, Tech. Rep., 2017.

[144] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[145] Q. Huang and Z. Liao, "A convolutional temporal encoder for video caption generation," Tech. Rep.

[146] Y. Liu, X. Li, and Z. Shi, "Video captioning with listwise supervision," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

[147] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4203–4212.

[148] 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. (ICCV)*, Oct. 2017, pp. 706–715.

[149] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3185–3194.

[150] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, and H. Tao Shen, "Hierarchical LSTM with adjusted temporal attention for video captioning," 2017, *arXiv:1706.01231*.

[151] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention LSTM networks for video captioning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 537–545.

[152] X. Li, B. Zhao, and X. Lu, "MAM-RNN: Multi-level attention model based RNN for video captioning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2208–2214.

[153] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1141–1150.

[154] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3135–3144.

[155] T.-H. Chen, K.-H. Zeng, W.-T. Hsu, and M. Sun, "Video captioning via sentence augmentation and spatio-temporal attention," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 269–286.

[156] Z. Yang, Y. Han, and Z. Wang, "Catching the temporal regions-of-interest for video captioning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 146–153.

[157] A.-A. Liu, N. Xu, Y. Wong, J. Li, Y.-T. Su, and M. Kankanhalli, "Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language," *Comput. Vis. Image Understand.*, vol. 163, pp. 113–125, Oct. 2017.

[158] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim, "Supervising neural attention models for video captioning by human gaze data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6119–6127.

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

[160] S. Chen, J. Chen, Q. Jin, and A. Hauptmann, "Video captioning with guidance of multimodal latent topics," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1838–1846.

[161] 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.

[162] 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.

[163] Y. Tu, X. Zhang, B. Liu, and C. Yan, "Video description with spatial–temporal attention," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1014–1022.

[164] 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.

[165] S. Li, J. Zhang, Q. Guo, J. Lei, and D. Tu, "Generating video description with long-short term memory," in *Proc. Int. Conf. Image, Vis. Comput. (ICIVC)*, Aug. 2016, pp. 73–78.

[166] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko, "Multimodal video description," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1092–1096.

[167] C. Zhang and Y. Tian, "Automatic video description generation via LSTM with joint two-stream encoding," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2924–2929.

[168] Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional long-short term memory for video description," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 436–440.

[169] W. Yue, W. Xiaojie, and M. Yuzhao, "First-feed LSTM model for video description," *J. China Universities Posts Telecommun.*, vol. 23, no. 3, pp. 89–93, Jun. 2016.

[170] R. Shetty and J. Laaksonen, "Frame- and segment-level features and candidate pool evaluation for video caption generation," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1073–1076.

[171] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen, "Attention-based LSTM with semantic consistency for videos captioning," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 357–361.

[172] Y. Liu and Z. Shi, "Boosting video description generation by explicitly translating from frame-level captions," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 631–634.

[173] N. Laokulrat, S. Phan, N. Nishida, R. Shu, Y. Ehara, N. Okazaki, Y. Miyao, and H. Nakayama, "Generating video description using sequence-to-sequence model with temporal attention," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Pape*, 2016, pp. 44–52.

[174] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4584–4593.
[175] 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.
[176] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1029–1038.
[177] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," 2014, *arXiv:1412.4729*.
[178] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.
[179] G. Li, S. Ma, and Y. Han, "Summarization-based video caption via deep neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1191–1194.
[180] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 190–200.
[181] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.
[182] S. Pini, M. Cornia, F. Bolelli, L. Baraldi, and R. Cucchiara, "M-VAD names: A dataset for video captioning with naming," *Multimedia Tools Appl.*, vol. 78, pp. 14007–14027, May 2019.
[183] 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.
[184] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 510–526.

TANZILA KEHKASHAN received the Master of Science degree in computer science from the University of Central Punjab, Lahore, Pakistan. She is currently pursuing the Ph.D. degree with Universiti Teknologi Malaysia (UTM). She is also a Lecturer with the Faculty of Computer Science, University of Lahore, Pakistan. Along with a rich professional journey, she has been contributing significantly to pursuing the Ph.D. degree. She is also an esteemed member of the Virtual, Visualization, and Vision Research Group (UTM VicubeLab). Her passion lies at visual computing, the intersection of computer vision, and natural language processing. Her research efforts are equally commendable, with her published work in renowned journals and conference proceedings. Her commitment to advancing knowledge is further demonstrated by her role as a supervisor for master's theses and final-year projects. Her research interests include deep learning, image/video analysis, medical imaging, and language modeling. Her combined dedication to academia, research, and professional practice reflects her passion for the field of computer science.

ABDULLAH ALSAEEDI received the B.Sc. degree in computer science from the College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, in 2008, the M.Sc. degree in advanced software engineering from the Department of Computer Science, The University of Sheffield, Sheffield, U.K., in 2011, and the Ph.D. degree in computer science from The University of Sheffield, in 2016. He is currently an Associate Professor with the Computer Science Department, Taibah University. His research interests include software engineering, software model inference, grammar inference, and machine learning.

WAEL M. S. YAFOOZ received the bachelor's degree in computer science in Egypt, in 2002, and the Master of Science and Ph.D. degrees in computer science from the University of MARA Technology (UiTM), Malaysia, in 2010 and 2014, respectively. He is currently an Associate Professor with the Computer Science Department, Taibah University, Saudi Arabia. He was awarded many gold and silver medals for his contribution to a local and international expo of innovation and invention in the area of computer science. Besides, he was awarded the Excellent Research Award from UiTM. He served as a member for various committees in many international conferences. Additionally, he chaired IEEE international conferences in Malaysia and China. Besides, he is a volunteer reviewer with different peer-reviewed journals. Moreover, he has supervised number of students at the master's and Ph.D. levels. Furthermore, he delivered and conducted many workshops in the research area and practical courses in data management, visualization, and curriculum design in area of computer science. He was invited as a speaker in many international conferences held in Bangladesh, Thailand, India, China, and Russia. His research interests include data mining, machine learning, deep learning, natural language processing, social network analytics, and data management.

NOR AZMAN ISMAIL received the Master of Information Technology degree from the National University of Malaysia and the Ph.D. degree in human–computer interaction from Loughborough University, U.K. He is currently an Associate Professor. His scholarly endeavors focus on user experience (UX), multimodal interaction, computer vision, web mining, and machine learning. With a distinguished career with Universiti Teknologi Malaysia (UTM), he has was the Deputy Dean overseeing research and innovation with the Faculty of Computing and previously as the Associate Chair responsible for both research and academic personnel with the School of Computing. Beyond academia, he has contributed significantly in administrative roles, including the University Web Director and the Deputy Director of Corporate Affairs. His pivotal role as the Head of the VicubeLab Research Group and a Research Fellow with the Media and Game Innovation Centre of Excellence (MaGIC-X) underscore his expertise and commitment to advancing research and innovation in his field.

ARAFAT AL-DHAQM received the B.Sc. degree in computer science from the University of Technology, Iraq, the M.Sc. degree in information security, and the Ph.D. degree in computer science from University Technology Malaysia (UTM). He is currently a Senior Lecturer and a Cybersecurity Researcher with the Faculty of Computing, UTM. He is also a member of the Information Assurance and Security Research Group (IASRG). He has a solid foundation in information security, digital forensics, information security governance, and risk management. He was trained by Cyber Security Malaysia (CSM) as a Certified Digital Forensic Investigator and the Certified Information Security Awareness Manager (CISM).