A Survey on Multi-modal Summarization

ANUBHAV JANGRA, Department of Computer Science, Indian Institute of Technology Patna, India
SOURAJIT MUKHERJEE, Department of Mathematics, Indian Institute of Technology Patna, India
ADAM JATOWT, Department of Informatics & DiSC, University of Innsbruck, Austria
SRIPARNA SAHA, Department of Computer Science, Indian Institute of Technology Patna, India
MOHAMMAD HASANUZZAMAN, Department of Computer Science, Cork Institute of Technology, Ireland

The new era of technology has brought us to the point where it is convenient for people to share their opinions over an abundance of platforms. These platforms have a provision for the users to express themselves in multiple forms of representations, including text, images, videos, and audio. This, however, makes it difficult for users to obtain all the key information about a topic, making the task of automatic multi-modal summarization (MMS) essential. In this article, we present a comprehensive survey of the existing research in the area of MMS, covering various modalities such as text, image, audio, and video. Apart from highlighting the different evaluation metrics and datasets used for the MMS task, our work also discusses the current challenges and future directions in this field.

CCS Concepts: • **Information systems** → *Similarity measures*; *Specialized information retrieval*; *Combination, fusion and federated search*; *Language models*; *Top-k retrieval in databases*; *Speech / audio search*; *Video search*; *Image search*; *Retrieval efficiency*; *Summarization*; *Information extraction*; • **Computing methodologies** → *Neural networks*; *Supervised learning*; *Unsupervised learning*; *Natural language processing*; *Information extraction*;

Additional Key Words and Phrases: Summarization, multi-modal content processing, neural networks

ACM Reference format:
Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A Survey on Multi-modal Summarization. *ACM Comput. Surv.* 55, 13s, Article 296 (July 2023), 36 pages.
https://doi.org/10.1145/3584700

1 INTRODUCTION

Every day, the Internet is flooded with tons of new information coming from multiple sources. Due to the technological advancements, people can now share information in multiple formats with various modes of communication to be used at their disposal. This alarmingly increasing

Authors' addresses: A. Jangra, Department of Computer Science, Indian Institute of Technology Patna, Patna, Bihar, India, 801106; email: anubhav0603@gmail.com; S. Mukherjee, Department of Mathematics, Indian Institute of Technology Patna, Patna, Bihar, India; email: mailsourajit25@gmail.com; A. Jatowt, Department of Informatics & DiSC, University of Innsbruck, Innsbruck, Austria; email: jatowt@acm.org; S. Saha, Department of Computer Science, Indian Institute of Technology Patna, Patna, Bihar, India; email: sriparna.saha@gmail.com; M. Hasanuzzaman, Department of Computer Science, Cork Institute of Technology, Bishopstown, Cork, Ireland; email: hasanuzzaman.im@gmail.com.

0360-0300/2023/07-ART296 $15.00
https://doi.org/10.1145/3584700

amount of content on the Internet makes it difficult for the users to receive useful information from the torrent of sources, necessitating research on the task of **multi-modal summarization (MMS)**. Various studies have shown that including multi-modal data as input can indeed help improve the summary quality [67, 88]. Zhu et al. [194] claimed that having a pictorial summary can improve the user satisfaction by 12.4%, on average, over a plain text summary. The fact that nearly every content sharing platform has a provision to accompany an opinion or fact in multiple media forms, and every mobile phone has the feature to deliver that kind of facility are indicative of the superiority of a multi-modal means of communication in terms of ease in conveying and understanding information.

Information in the form of multi-modal inputs has been leveraged in many tasks other than summarization including multi-modal machine translation [14, 31, 32, 58, 160], multi-modal movement prediction [26, 82, 177], multi-modal question answering [155], multi-modal lexico-semantic classification [71], multi-modal keyword extraction [174], product classification in e-commerce [186], multi-modal interactive artificial intelligence frameworks [80], multi-modal emoji prediction [7, 25], multi-modal frame identification [13], multi-modal financial risk forecasting [90, 148], multi-modal sentiment analysis [117, 138, 179], multi-modal named identity recognition [3, 115, 116, 161, 183, 189], multi-modal video description generation [55, 56, 135], multi-modal product title compression [107], and multi-modal biometric authentication [42, 62, 157]. The shear number of application possibilities for multi-modal information processing and retrieval tasks are quite impressive. Research on multi-modality can also be utilized in other closely related research problems such as image-captioning [18, 19], image-to-image translation [59], seismic pavement testing [139], aesthetic assessment [84, 101, 188], and visual question-answering [78].

Text summarization is one of the oldest problems in the fields of **natural language processing (NLP)** and **information retrieval (IR)**, that has attracted various researchers due to its challenging nature and potential for many applications. Research on text summarization can be traced back more than six decades [104]. The NLP and IR community have tackled research in text summarization for multiple applications by developing myriad of techniques and model architectures [21, 66, 102, 151]. As an extension to this, the problem of multi-modal summarization adds another angle by incorporating visual and aural aspects into the mix, making the task more challenging and interesting to tackle. This extension of incorporating multiple modalities into a summarization problem expands the breadth of the problem, leading to wider application range for the task. In recent years, multi-modal summarization has experienced many new developments, including release of new datasets, advancements in techniques to tackle the MMS task, as well as proposals of more appropriate evaluation metrics. The idea of multi-modal summarization is a rather flexible one, embracing a broad range of possibilities for the input and output modalities and also making it difficult to apprehend existing works on the MMS task with knowledge of uni-modal summarization techniques alone. This necessitates a survey on multi-modal summarization.

The MMS task, just like any uni-modal summarization task, is a demanding one, and existence of multiple correct solutions makes it very challenging. Humans creating a multi-modal summary have to use their prior understanding and external knowledge to produce the content. Establishing computer systems to mimic this behavior becomes difficult given their inherent lack of human perception and knowledge, making the problem of automatic multi-modal summarization a non-trivial but interesting task.

Quite a few survey papers were written for uni-modal summarization tasks, including surveys on text summarization [45, 49, 64, 120, 165, 181] and video summarization [8, 61, 81, 114, 149] and a few survey papers covering multi-modal research [4, 6, 63, 134, 150, 158]. However, to the best of our knowledge, we are the first to present a survey on multi-modal summarization. The closest work to ours is the work on multi-dimensional summarization by Zhuge [197], who proposes

the method for summarization of things in cyber-physical society through a multi-dimensional lens of semantic computing. However, our survey is distinct from that work, as Zhuge [197] focuses on how understanding human behavior, psychology, and advances in cognitive sciences can help to improve the current summarization systems in the emerging cyber-physical society; while, in this manuscript, we mostly focus on the direct applications and techniques adopted by the research community to tackle the MMS task. Through this manuscript, we unify and systematize the information presented in related works, including the datasets, methodology, and evaluation techniques. With this survey, we aim to assist researchers familiarize with various techniques and resources available to proceed with research in the area of multi-modal summarization.

The rest of the article is structured as follows: We formally define the MMS task in Section 2. In Section 3, we provide an extensive organization of existing works. In Section 4, we give an overview about the techniques used for the task of MMS. In Section 5, we introduce the datasets available for the MMS task and evaluation techniques devised for the evaluation of multi-modal summaries, respectively. We discuss about possibilities of future work in Section 7 and conclude our article in Section 8.

2 MULTI-MODAL SUMMARIZATION TASK

In this section, we formally define what classifies as a multi-modal summarization task. Before formalizing the multi-modal summarization, we broadly define the term summarization.[1] According to Wikipedia,[2] automatic summarization is "*the process of shortening a set of data computationally, to create an abstract that represents the most important or relevant information within the original content.*" Formally, summarization is the process of obtaining the set $X_{sum} = f(D)$ such that $length(X_{sum}) \leq length(D)$, where X_{sum} is the output summary, D is the input data, and function $f(.)$ is the summarization function.

The multi-modal summarization task can be defined as a summarization task that takes more than one mode of information representation (termed as modality) as input and depends on information sharing across different modalities to generate the final summary. Mathematically speaking, when the input dataset D can be broken down into several partially disjoint sets of different modality information $\{M_1 \bigcup M_2 \bigcup \cdots \bigcup M_n\}$, where $n \geq 2$ and $\exists$ several pairs of (M_i, M_j) for $(i, j) \in \{1, \ldots, n\}$ such that the shared latent information between (M_i, M_j) is not $\varnothing$, then the task of obtaining the set $X_{sum} = f(D)$ is known as multi-modal summarization.[3] If $n' \geq 2$ for $X_{sum} = \{M'_1 \bigcup M'_2 \bigcup \cdots \bigcup M'_{n'}\}$, then the output summary is multi-modal; otherwise, the output is a uni-modal summary.

In this survey, we mainly focus on recent works that have natural language as the *central modality*,[4] where a *central modality* (or *key modality*) is selected according to the intuition: *"For any information processing task in multi-modal scenarios, including content summarization, amongst all the modalities, there is often a preferable mode of representation based on the significance and ability to fulfill the task"* [69]. Other modalities that aid the central modality to convey information are termed as *adjacent modalities*.

[1] In this article, summarization stands for automatic summarization unless specified otherwise.
[2] https://en.wikipedia.org/wiki/Automatic_summarization.
[3] The reason for restricting $n \geq 2$ for the task definition is a limitation of current techniques, which are unable to successfully generate modalities other than text for multi-modal summarization. Even though there have been some recent breakthroughs in text-to-image generation (like Open AI's DALL-E [136]) and text-to-speech synthesis (like Google's Duplex [85]), they still lack the level of integrity and robustness to be used in a real-world application like MMS.
[4] We believe that the MMS models that have video as the *central modality* tend to be closely related to the task of video summarization.

Various aspects of multi-modal summarization: Literature has explored the MMS task for myriad reasons and motives and doing so has led to different challenges and variants of the task. Some of the most prominent and interesting ones are discussed below:

- **Combined complementary-supplementary multi-modal summarization task (CCS-MMS) [69]:** Jangra et al. [69] proposed the CCS-MMS task of generating a multi-modal summary that considers text as the central modality and images, audio, and videos as the adjacent modality. The task is to generate the multi-modal summary such that it consists of both supplementary and complementary enhancements, which are defined as follows:
  - **Supplementary enhancement:** When the adjacent modalities reinforce the facts and ideas presented in the central modality, the adjacent modalities are termed as *supplementary enhancements*.
  - **Complementary enhancement:** When the adjacent modalities complete the information by providing additional but relevant information that is not covered in the central modality, the adjacent modalities are termed as *complementary enhancements*.
- **Summarization objectives:** We can distinguish prior work based on summarization objectives they have used. For instance, Li et al. [88] use weighted sum of three submodular objective functions to create an extractive text summarization system that is guided by multi-modal inputs. The chosen submodular functions are salience of input text, image information covered by text summary, and non-redundancy in input text. Jangra et al. [67] use a single objective function for an ILP setup, which is the weighted average of uni-modal salience, and cross-modal correspondence. Jangra et al. [68] propose two different sets of multi-modal objectives for the task of extractive multi-modal summary generation: (a) summarization-based objective and (b) clustering-based objectives. For summarization-based objectives, they use the following three objectives: (i) Salience(txt) / Redundancy(txt), (ii) Salience(img)/Redundancy(img), and (iii) cross-modal correspondence; while for clustering-based objectives, they use PBM [123], a popular cluster validity index (a function of cluster compactness and separation) to evaluate the uni-modal clusters of image and text, giving the following set of objectives: PBM(txt), PBM(img), and cross-modal correspondence. Almost all the neural networks based multi-modal summarization frameworks [16, 17, 87, 124], however, use the standard negative log-likelihood function over the output vocabulary as the training objective. Some works also use textual and visual coverage loss to prevent over-attending the input as well [87, 194].
- **Multi-modal social media event summarization:** Various works have been conducted on the social media data that consists of opinions and experiences of a diverse set of population. Tiwari et al. [167] propose the problem of summarizing asynchronous information from multiple social media platforms such as Twitter, Instagram, and Flickr to generate a summary of events that is widely covered by users of these platforms extensively. Bian et al. [10] propose multi-modal summarization of trending topics in microblogs. They use Sina Weibo[5] microblogs for the experimentation, which is a very popular microblogging platform in China. Qian et al. [132] use the Weibo platform information to summarize disaster events such as train crashes and earthquakes.

3 ORGANIZATION OF EXISTING WORK

Different attempts have been made to solve the MMS task, and thus it is important to categorize the existing works to get a better understanding of the task. We categorize the prior works into three broad categories, depending upon encoding the input, the model architecture, and decoding the

[5] http://www.weibo.com/.

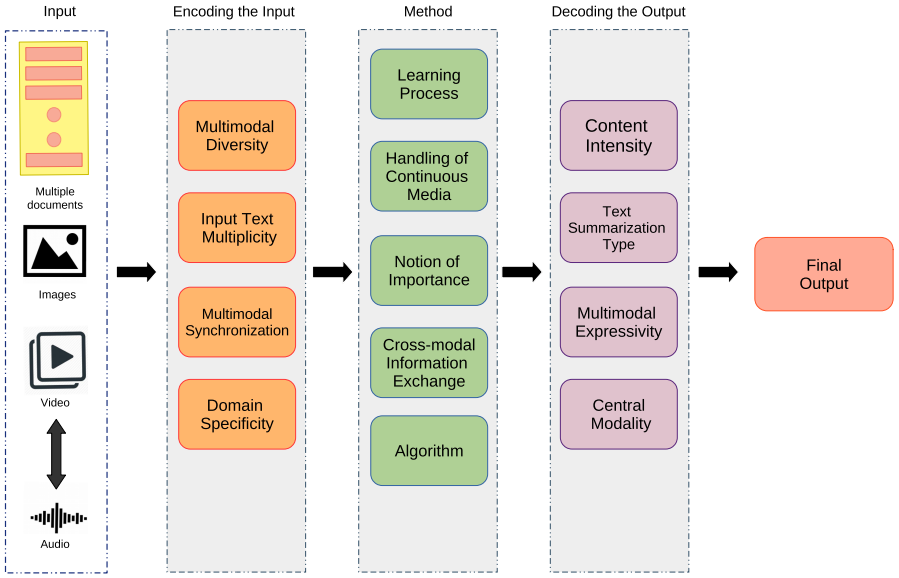


Fig. 1. Generic multimodal summarization model flow diagram based on the defined taxonomy. Here, the boxes at different stages of the flow diagram highlight the various types of factors that need to be taken care of during that stage. "Orange" box for "Encoding the Input" stage, "Green" box for "Method" stage, and "Purple" for "Decoding the output" stage. Based on these factors, we have defined our taxonomy.

output. We have also illustrated these categorizations through a generic model diagram in Figure 1. A detailed pictorial representation of the taxonomy is shown in Figure 2, and a comprehensive study is provided in Table 2 (note that if some classifications are not marked in the table, then either the information about that category was not present, or is not applicable.).

3.1 On the Basis of Encoding the Input

A multi-modal summarization task is highly driven by the kind of input it is given. Due to this dependency on diverse input modalities, the feature extraction and encoding strategies also differ for different multi-modal summarization systems. Existing works can be distinguished from others on the basis of the type of input and its encoding strategy in the following categories:

Multi-modal Diversity (MMD): Different combinations of input (text, image, video, and audio) involve different preprocessing and encoding strategies. We can classify the existing works depending on the combination of modalities in which the input is represented. Various combinations within the input modalities such as *text-image* [16, 87, 194], *text-video* [43, 92], *audio-video*[6] [35, 38], and *text-image-audio-video* [67–69, 88, 171] have been explored in the literature of MMS. The different feature extraction strategies for individual modalities are described in Section 3.1.1.

Input Text Multiplicity (ITM): Since a major focus of this survey is on MMS tasks with text as the *central modality*, the number of text documents in input can also be one way of categorizing the related works. Depending upon whether the textual input is single-document [17, 87, 194] or multi-document [67–69, 88], the input preprocessing and the overall summarization strategies might differ. Having multiple documents makes the task a lot more challenging, since the degree

[6]Note that *audio-video* and *text-audio-video* works are grouped together, since in most of the existing works, automatic speech transcription is performed to obtain the textual modality part of data in the pre-processing step.

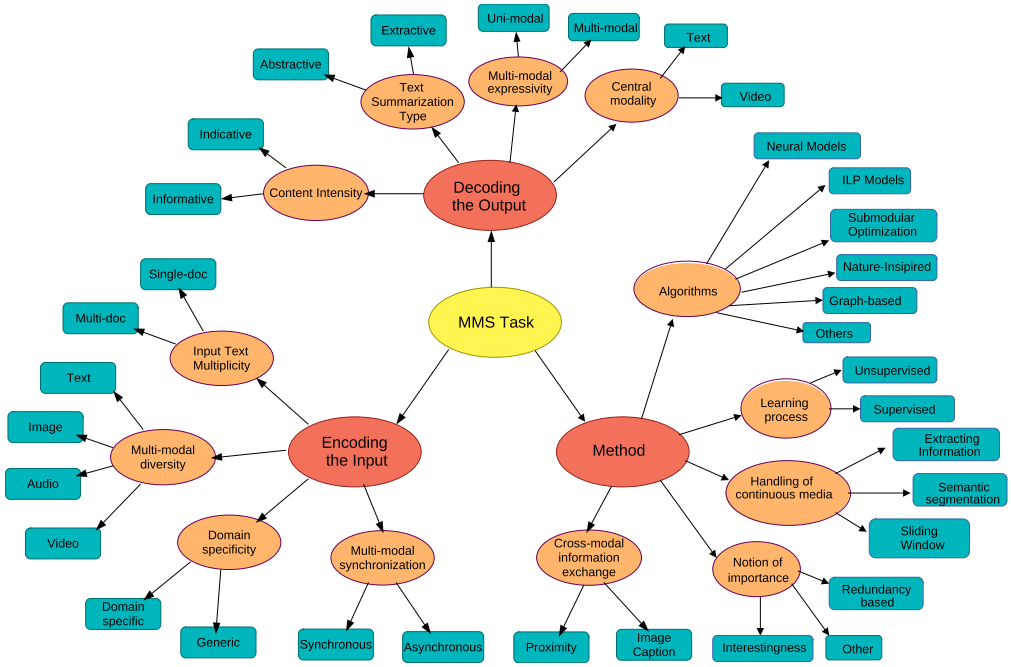


Fig. 2. Visual representation of proposed taxonomy. The dark-orange nodes coming out of the root (in yellow) represent the segregation based on input, output, and adopted methodology, while the light-orange nodes following them represent the respective characteristics of the research work on which the works can be distinguished. The teal-colored rectangles in the leaf denote the various categories of each such characteristic.

of redundant information in input becomes a lot more prominent, making the data somewhat more noisy [105].

Multi-modal Synchronization (MMSy):[7] Synchronization refers to the interaction of two or more things at the same time or rate. For multi-modal summarization, having a synchronized input indicates that the multiple modalities have a coordination in terms of information flow, making them convey information in unison. We then classify input as synchronous [35, 38] and asynchronous [67–69, 88, 168].

Domain Specificity (DS): Domain can be defined as the specific area of cognition that is covered by any data, and depending upon the extent of domain coverage, we can classify works as *domain-specific* or *generic*. The approach to summarize a *domain-specific* input can differ from the *generic* input greatly, since feature extraction in the former can be very particular in nature while not so in the latter, impacting the overall techniques immensely. Most of the news summarization tasks [16, 68, 69, 88, 194] are *generic* in nature, since news covers information about almost all the domains; whereas movie summarization [38], sports event summarization for tennis [168] and soccer [146], meetings recording summarization [35], tutorial summarization [93], social media event summarization [167] are examples of *domain-specific* tasks.

3.1.1 Feature Extraction Strategies. In a multi-modal setting, pre-processing and feature extraction becomes vital step, since it involves extracting features from different modalities. Each input

[7]Note that the term synchronization is mostly used when there is a continuous media in consideration.

Table 1. Comprehensive List of Works that Use Specific Pre-trained Deep Learning Frameworks Used to Generate Image Embeddings

Pre-trained network	Works using this framework
VGGNet [154]	Chen and Zhuge [16, 17], Jangra et al. [67, 68, 69], Li et al. [87, 88], Modani et al. [113], Zhu et al. [194, 195]
ResNet [52]	Fu et al. [43], Li et al. [86, 92]
GoogleNet [163]	Sanabria et al. [146]

modality has been dealt with using modal-specific feature-extraction techniques. Even though some works tend to learn the semantic representation of data using their own proposed models, nearly all follow the same steps for feature extraction. Since the related works have different sets of input modalities, we describe feature extraction techniques for each modality individually.

Text: Traditionally, before the era of deep learning, **Term Frequency-Document Inverse Frequency (TF-IDF)** [145] was used to identify relevant text segments [35, 38, 168]. Due to significant advancements in feature extraction, almost all the MMS tasks in the past five years either use pre-trained embeddings such as word2vec [110] or GloVe [128]. These pre-trained embeddings utilize the fact that the semantic information of a word is related to its contextual neighbors for training. Some works also train similar embeddings on their own datasets [194, 195] (refer to *Feature Extraction* in Section 4.1.1). Some works also adopt different pre-processing steps, depending upon the task specifications. For example, Tiwari et al. [167] applied a normalizer to handle the concept of expressive lengthening dealing with microblog datasets. Even though current MMS systems have not yet adopted them, it is worth mentioning Transformer-based word representations [173] like BERT, and so on, that have achieved state-of-the-art performance in the vast majority of NLP and vision tasks. This achievement can be credited to their fast training due to parallelization and ability to pre-train the language models on unlabelled corpora. We even have multi-lingual embeddings like LabSE [41], and multi-modal text-image embeddings such as UNITER [22], VilBERT [103], VisualBERT [91], Pixel-BERT [60], and so on.

Images: Images, unlike text, are non-sequential and have a two-dimensional contextual span. **Convolutional neural network (CNN)**-based deep neural network models have proven to be very promising in feature extraction tasks, but training these models requires large datasets, making it difficult to train features on MMS datasets. Hence, most of the existing works use pre-trained networks (e.g., ResNet [52], VGGNet [154], GoogleNet [163]) trained on large image classification datasets like ImageNet [28]. The technique of extracting local features (containing information about a confined patch of image) along with global features has shown promise in the MMS task as well [194]. A detailed list of frameworks that use pre-trained deep learning networks can be found in Table 1. Tiwari et al. [167] use **Speeded-Up Robust Features (SURF)** for each image, following a bag-of-word approach to creating a visual vocabulary. Chen and Zhuge [17] handle images by first extracting the **Scale Invariant Feature Transform (SIFT)** features. These SIFT features are fed to a hierarchical quantization module [131] to obtain a 10,000-dimensional bag of the visual histogram. Having been inspired by the success of self-attention and Transformers [173] in effectively modeling textual sequences, researchers in computer vision have adopted the techniques such as self-attention, unsupervised pre-training, parallelizability of transformer architecture, and so on, to better model the image representations.[8] To adopt the self-attention layer dedicated to

[8]The readers are encouraged to read the extensive survey provided by Khan et al. [75].

text sequences, Parmar et al. [125] proposed a framework that restricts the self-attention to the local neighborhoods, thus significantly increasing the size of images that the model can process, despite maintaining larger receptive fields per layer than a CNN framework. Dosovitskiy et al. [30] illustrated that usage of self-attention in conjunction with CNNs is not required, and a pure transformer applied to the sequence of image patches can also perform well on image classification tasks. Touvron et al. [169] developed and optimized deep image transformer frameworks that do not saturate early with more depth.

To the best of our knowledge, none of the existing multi-modal summarization works use image transformers to encode the images. Since these large-scale models have a lot more capability to store more learned patterns from large-scale datasets due to the huge parameter space, they are bound to improve the overall summarization process by aiding in better image understanding.

Audio and video: Audio and video are usually present together as a single synchronized continuous media, and hence, we discuss the pre-processing techniques used to extract features from them simultaneously. Continuous media has been processed in many diverse ways. Since audios and videos are susceptible to noise, it becomes of utmost importance to detect relevant segments before proceeding to the training phase.[9] While some works have adopted a naïve sliding window approach, making equal length cuts, and further experimenting on these segments [35], quite a few have done a modal conversion, changing the information media using automatic speech transcription to generate speech transcriptions and extracting key-frames from video using techniques like boundary shot-detection [67–69, 88, 168]. Some works have also taken into account the nature of the dataset and performed semantic segmentation, getting better segment slices. For example, Tjondronegoro et al. [168] worked on a tennis dataset and used the information that the umpire requires the audience to remain quiet during the match point, performing segmentation consisting of a segment that begins with low audio activity followed by high audio energy levels as a result of the cheering and the commentary. If the audio and video are converted into another modality, then their pre-processing follows the same procedure as the new modalities, whereas, in the case of segmentation, various metrics such as acoustic confidence, audio magnitude, sound localization for audio, motion detection, and Spatio-temporal features driven by intensity, color, and orientation for video have been explored to determine the salience and relevance of segments depending upon the task at hand [35, 38, 88].

Cross-modal correspondence: Although the majority of works train their own shared embedding space for multiple modalities using the information from the target datasets [87, 93, 194], quite a few works [67–69, 88, 113] also tend to use pre-trained neural network models [72, 176] trained on the image-caption datasets such as Pascal1k [137], Flickr8k [54], Flick30k [182], and so on, to leverage the information overlap amongst different modalities. This becomes a necessity for small datasets that are mostly used for extractive summarization. However, even these pre-trained models cannot process raw data, and hence the text and image inputs are first pre-processed to desired embedding formats and then are fed to these models with pre-trained weights. For example, Wang et al. [176] required a 6,000-dimensional sentence vector and 4,096-dimensional image vector generated by applying **Principal Component Analysis (PCA)** [127] to the 18,000-dimensional output from the **Hybrid Gaussian Laplacian mixture model (HGLMM)** [83] and extracting the weights from the final fully connected layer, $fc7$, from VGGNet [154], respectively. In recent years, various Transformer-based [173] models have also been developed to correlate semantic information across textual and visual modalities. These BERT [29]-inspired models include ViLBERT [103],

[9]Note that some deep neural models such as Fu et al. [43] or Li et al. [92] prefer to encode individual frames using CNNs and then use trainable RNNs to encode temporal information in videos. This CNN-RNN framework is not part of pre-processing, but instead, it belongs to the main model, since these layers are also affected during training.

VisualBERT [91], VideoBERT [162], VLP [193], Pixel-BERT [60], and so on, to name a few. There has also been some video-text representation learning, such as References [96] and [126], that can be used to summarize multi-modal content with continuous modalities. However, none of the recent works on multi-modal summarization has utilized these transformer-based techniques in their system pipelines.

Domain-specific techniques: Most of the systems proposed to solve the problem of multi-modal summarization are generic and can be adapted to other domains and problem statements as well. But, there do exist some works that benefit from the external knowledge of particular domains and problem settings to create better-performing systems. For instance, Tjondronegoro et al. [168] utilize the fact that in tennis, the umpire always requires spectators to be silent before a serve, until the end of the point. The authors also pointed out that the end of the point is usually marked by a loud cheer from the supporters of the players in the audience. They used this fact to perform smooth segmentation of tennis clips using audio energy levels to indicate the start and end positions of a segment. Similar to this, Sanabria et al. [146] utilized atomic events in a game of soccer such as a pass, goal, dribble, and so on, to segment the video, which is later connected together to generate the summary. Other than sports, such domain-specific solutions have also been adopted in other domains. For example, Erol et al. [35], when summarizing meeting recordings of a conference room, seek out some visual activity such as "someone entering the room" or "someone standing up to write something on a whiteboard" to detect some event likely to contain relevant information. In a different domain setting, people have benefited from other data pre-processing strategies; for instance, Li et al. [86] extract various key aspects of products such as "environmentally friendly refrigerators" or "energy efficient freezers" to generate a captivating summary for Chinese e-commerce products.

3.2 On the Basis of Method

A lot of various approaches have been developed to solve the MMS task, and we can organize the existing works on the basis of proposed methodologies as follows:

Learning process (LP): A lot of work has been done in both supervised learning [17, 87, 93, 194, 195] and unsupervised learning [35, 38, 67–69, 88]. It can be observed that a large fraction of supervised techniques adopt deep neural networks to tackle the problem [16, 87, 93, 194], whereas in unsupervised techniques a large diversity of techniques has been adopted, including deep neural networks [17], integer linear programming [67], differential evolution [68, 69], submodular optimization [88], and so on.

Handling of continuous media (HCM): We can also distinguish between works depending upon how the proposed models handle continuous media (audio and video in this case). There are three broad distinctions possible: (a) *extracting-information*, where the model extracts information from continuous media to get a discrete representation [67–69, 88]; (b) *semantic-segmentation*, where a logical technique is proposed to slice out the continuous media [38, 39, 168]; and (c) *sliding window*, when a naïve fixed window-based modeling is performed [35].

Notion of importance (NI): One of the most significant distinctions would be the notion of importance used to generate the final summary. A diverse set of objectives ranging from interestingness [168], redundancy [88], cluster validity index [68], acoustic energy/visual illumination [38, 39], and social popularity [141] has been explored in attempt to solve the MMS task.

Cross-modal information exchange (CIE): The most important part of a MMS model is the ability to extract and share information across multiple modalities. Most of the works either adopt a proximity-based approach [35, 178], a pre-trained model on image-caption pairs-based corpora

for information overlap [67–69, 88], or learn the semantic overlap over uni-modal embeddings [16, 17, 87, 194, 195].

Algorithms (A): The algorithm for the multimodal summarization task varies from traditional multiobjective optimization strategies to modern deep learning-based approaches. We can classify the existing works based on the different algorithms as **Neural models (NN), Integer Linear Programming based models (ILP), Submodular Optimization-based models (SO), Nature-Inspired Algorithm based models (NIA), Graph-based models (G)**, and **other algorithms (Oth)**. We have discussed these different methods in detail in Section 4.1. The other algorithms comprise different clustering-based, LDA [12]-based, and audio-video analysis-based techniques, which were earlier used for performing multimodal summarization.

3.3 On the Basis of Decoding the Output

The summarization objective decides the desired type of output. For different summarization objectives, the type of output and decoding method vary. Depending on the type of output and the decoding method, we can categorize the existing works on the following basis:

Content intensity (CI): The degree to which an output summary elaborates on a concept can hugely impact the overall modeling. The output summary can either be *informative*, having detailed information about the input topic [93, 180], or *indicative*, only hinting at the most relevant information [16, 194].

Text Summarization Type (TST): The most widely discussed distinction for text summarization works is the distinction of *extractive* vs. *abstractive*. Abstractive summarization systems generally use a beam search or greedy search mechanism for decoding the output summary; while extractive systems during decoding use some scoring mechanism to identify the salient, non-redundant, and readable elements from the input for the final output. Depending on the nature of an output text summary, we can also classify the works in MMS tasks (containing text in the output) into *extractive MMS* [17, 67–69, 88] and *abstractive MMS* [16, 87, 194, 195].[10]

Multi-modal expressivity (MME): Whether the output is *uni-modal* (comprising one modality) [17, 38, 87, 88, 93] or *multi-modal* (comprising multiple modalities) [16, 67–69, 168, 194, 195] is a major classification for the existing work. Mostly the systems producing multimodal output involve some post-processing steps for selecting the final output elements from the non-central modalities.

Central modality (CM): Based on *central-modality* (defined in Section 2), existing works can also be distinguished depending on the base modality around which the final output, as well as the model, are formulated. A large portion of the prior work adopts either a text-centric approach [16, 67, 69, 87, 88, 93, 194] or a video-centric[11] approach [35, 38, 141, 168]. A few of the decoding methods followed popularly in neural models have been discussed in detail in Section 4.1.1.

4 OVERVIEW OF METHODS

A lot of works have attempted to solve the MMS task using supervised and unsupervised techniques. In this section, we attempt to describe the MMS frameworks in a generalized manner, elucidating the nuances of different approaches. Since the variety of inputs, outputs, and techniques that were used span a large spectrum of possibilities, we describe each one individually. We have broken down this section into three stages: *pre-processing*, *main model*, and *post-processing*.

[10]Note that other modalities besides text have been so far subject to only extractive approaches in MMS researches.
[11]Here, audio is assumed to be a part of video, since in all the existing works, video and audio are synchronous to each other.

Table 2. Comprehensive Study of Existing Work Using the Proposed Taxonomy
(Refer to Section 3)

Papers	Input Based								Output Based								Method Based														
	MMD				ITM		MMSy	DS	CI		TST		MME		CM		LP		HCM			NI			CIE			A
	Text	Images	Audio	Video	Single-doc	Multi-doc	Sync	Async	Domain-specific	Generic	Informative	Indicative	Abstractive	Extractive	Uni-modal	Multi-modal	Text	Video	Unsupervised	Supervised	Extracting info	Semantic seg	Sliding window	Redundancy-based	Interestingness	Other	Image Caption	Proximity	Uni-modal embedding	Algorithms
Erol et al. [35]	✓		✓	✓	✓		✓		✓			✓				✓		✓	✓		✓					✓		✓		Oth
Tjondronegoro et al. [168]	✓		✓	✓	✓		✓		✓		✓					✓		✓	✓						✓			✓		Oth
UzZaman et al. [171]	✓	✓			✓			✓		✓	✓					✓			✓						✓					Oth
Evangelopoulos et al. [38]	✓	✓	✓	✓	✓		✓		✓			✓			✓	✓			✓		✓						✓			Oth
Li et al. [88]	✓	✓	✓	✓		✓	✓		✓			✓		✓					✓				✓							SO, G
Li et al. [87]	✓	✓			✓		✓		✓			✓	✓				✓		✓				✓						✓	NN
Zhu et al. [194]	✓	✓			✓		✓		✓			✓	✓			✓	✓		✓				✓						✓	NN
Chen and Zhuge [16]	✓	✓			✓		✓		✓			✓	✓			✓	✓		✓				✓						✓	NN
Libovický et al. [93]	✓		✓	✓	✓		✓		✓			✓	✓		✓			✓	✓				✓						✓	NN
Palaskar et al. [124]	✓		✓	✓	✓		✓		✓			✓	✓		✓			✓	✓				✓						✓	NN
Zhu et al. [195]	✓			✓	✓		✓		✓			✓	✓			✓	✓		✓				✓						✓	NN
Jangra et al. [67]	✓		✓	✓		✓		✓		✓			✓			✓						✓								ILP
Jangra et al. [68]	✓		✓	✓	✓		✓		✓			✓	✓			✓	✓		✓				✓			✓				NIA
Jangra et al. [69]	✓		✓	✓	✓		✓		✓			✓	✓			✓	✓		✓				✓			✓				NIA
Xu et al. [178]	✓		✓		✓		✓		✓		✓			✓		✓								✓	✓		✓			NIA
Sahuguet and Huet [141]	✓	✓	✓	✓	✓		✓		✓			✓	✓			✓	✓		✓					✓	✓		✓			Oth
Tiwari et al. [167]	✓	✓			✓		✓		✓			✓	✓			✓	✓		✓					✓			✓			SO
Bian et al. [10]	✓	✓			✓			✓		✓			✓			✓				✓				✓						Oth
Yan et al. [180]	✓	✓			✓		✓		✓		✓		✓			✓	✓		✓	✓					✓					G
Qian et al. [130]	✓	✓			✓		✓		✓		✓	✓	✓			✓	✓		✓				✓							Oth
Chen and Zhuge [17]	✓	✓			✓		✓		✓			✓	✓			✓	✓		✓				✓					✓		NN
Evangelopoulos et al. [39]	✓		✓	✓	✓		✓		✓			✓	✓			✓	✓		✓			✓			✓					Oth
Bian et al. [11]	✓	✓			✓			✓		✓			✓			✓				✓				✓						Oth
Fu et al. [43]	✓		✓	✓	✓		✓		✓			✓	✓			✓	✓		✓		✓			✓						NN
Li et al. [92]	✓			✓	✓		✓		✓			✓	✓			✓	✓		✓		✓		✓	✓						NN
Li et al. [86]	✓	✓			✓		✓		✓			✓	✓			✓	✓		✓					✓						NN
Modani et al. [113]	✓	✓			✓		✓		✓			✓	✓			✓	✓		✓					✓			✓			SO, G
Sanabria et al. [146]	✓		✓	✓	✓		✓		✓			✓	✓			✓	✓		✓		✓			✓			✓			NN

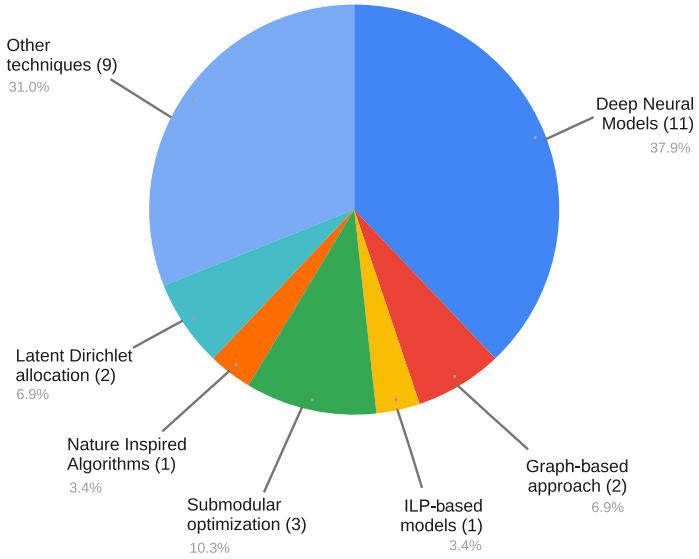


Fig. 3. Illustration of techniques adopted to solve the MMS task.

4.1 Main Model

A lot of different techniques have been adopted to perform the MMS task using extracted features. Figure 3 illustrates the analysis of techniques adopted by researchers to solve the MMS task. We have tried to cover almost all the recent architectures that mainly focus on text-centric output summaries. In the approaches that have text as the central modality, the adjacent modalities are treated as a supplement to the text summaries, often getting selected at the post-processing step (Section 4.2).

4.1.1 Neural Models. A few extractive summarization models [17] and almost all of the abstractive text summarization-based MMS architectures [16, 87, 93, 194, 195] use **Neural Networks (NN)** in one form or another. Obtaining annotated dataset with sufficient instances to train these supervised techniques is the most difficult step for any deep learning-based MMS framework. The existing datasets satisfying these conditions belong to news domain and have *text-image* type input (refer to datasets #4, #5, #6, #7, #19 in Table 3) or *text-audio-video* type input (refer to datasets #17, #18 in Table 3). All these frameworks utilize the effectiveness of seq2seq RNN models for language processing and generation and encoding temporal aspects in videos; CNN networks are also adopted to encode discrete visual information in form of images [16, 194] and video frames [43, 92]. All the neural models have an encoder-decoder architecture at their heart, having three key elements: (1) a *feature extraction module (encoder)*, (2) a *summary generation module (decoder)*, and (3) a *multi-modal fusion module*. Figure 4 describes a generic neural model to generate *text-image*[12] summaries for multi-modal input.

Feature Extraction (Encoder): Encoder is a generic term that entails both textual encoders as well as visual encoders. Various *encoders* have been explored to encode contextual information in textual modality, ranging from *sentence-level encoders* [87] to *hierarchical document level encoders* [16] with ***Long Short Term Memory (LSTM)*** *units* [53] or ***Gated Recurrent Units (GRU)*** [23] as

[12]We formulate *text-image* summaries in our generic model, since the existing neural models only output text [17, 43, 87] or text-image [16, 92, 194, 195] output.

Table 3. A Study on Datasets Available for Multi-modal Summarization

ID & Paper	Used In Paper	Input Modalities	Output Modalities	Data Statistics	Domain
#1: Li et al. [87] (2018)	[87]	T, I	TA	66,000 triplets (sentence, image, and summary)	News
#2: Zhu et al. [194](2018)*	[194, 195]	T, I	TA, I	313k documents, 2.0m images	News
#3: Chen and Zhuge [16] (2018)	[16]	T, I	TA, I	219k documents	News
#4: Xu et al. [178] (2013)	[178]	T, I	TE, I	8 topics (each containing 150+ documents)	News
#5: Bian et al. [10] (2013)	[10]	TC, I	TE, I	10 topics (127k microblogs and 48k images)	Social Media
#6: Bian et al. [11] (2014)	[11]	TC, I	TE, I	20 topics (310k documents, 114k images)	Social Media
#7: Li et al. [86] (2020)	[86]	TC, I	TA	1,375,453 instances from home appliances, clothing, and cases & bags categories	E-commerce
#8: Chen and Zhuge [17] (2018)	[17]	T, A	TE	-	News
#9: Tiwari et al. [167] (2018)	[167]	T, I, TM	TE, I	6 topics	Social Media
#10: Yan et al. [180] (2012)	[180]	T, I, TM	TE, I	4 topics (6k documents, 2k images)	News
#11: Qian et al. [130] (2019)	[130]	T, I, U	TE, I	12 topics (9.1m documents, 2.2m users, 15m images)	News (disasters)
#12: Tjondronegoro et al. [168] (2011)	[168]	T, A, V	TF	66 hrs video (33 matches), 1,250 articles related to Australian Open 2010 tennis tournament	Sports (Tennis)
#13: Sanabria et al. [147] (2018)*	[93]	T, A, V	TA	2,000 hrs video	Multiple domains
#14: Fu et al. [43] (2020)	[43]	T, A, V	TA	1970 articles from *Daily Mail* (avg. video length 81.96 secs), and 203 articles from CNN (avg. video length 368.19 secs)	News
#15: Li et al. [92] (2020)	[92]	T, A, V	TA, I	184,920 articles (Weibo) with avg. video duration 1 min, avg. article length 96.84 words, avg. summary length 11.19 words	News
#16: Sanabria et al. [146] (2019)	[146]	T, A, V	A, V	20 complete soccer games from 2017–2018 season of French Ligue 1	Sports (Soccer / Football)
#17: Evangelopoulos et al. [39] (2009)	[39]	T, A, V	A, V	3 movie segments (5–7 min each)	Movies
#18: Evangelopoulos et al. [38] (2013)	[38]	T, A, V	A, V	7 half-hour segments of movies	Movies
#19: Jangra et al. [67] (2020)	[67, 68]	T, I, A, V	TE, I, A, V	25 topics (500 documents, 151 images, 139 videos)	News
#20: Jangra et al. [69] (2021)*	[69]	T, I, A, V	TE, I, A, V	25 topics (contains complementary and supplementary multi-modal references)	News
#21: Li et al. [88] (2017)*	[88]	T, TC, I, A, V	TE	25 documents in English, 25 documents in Chinese	News

"T" stands for English text, "TC" stands for Chinese text, "TF" stands for text (template filling), "TE" stands for text (extractive), "TA" stands for text (abstractive), "I" stands for images, "V" stands for video, "A" stands for audio, "U" signifies user information, and "TM" denotes existence of temporal information about the data such as publication date. The "*" denotes publicly available datasets and the "-" denotes the unavailability of details.

the underlying RNN architecture. Most of the visual encoders do not train the parameter weights from scratch, but rather prefer to use CNN-based pre-trained embeddings (refer to Section 3.1.1). However, notably, to capture the contextual information of images, Chen and Zhuge [16] used a bi-directional GRU unit to encode information from multiple images (encoded using VGGNet [154]) into one context vector, which is a unique approach for discrete image inputs. However, this RNN-CNN-based encoding strategy is a very standard approach adopted to encoding video input. Fu et al. [43] and Li et al. [92], in their respective works, use pre-trained CNNs to encode individual frames and then feed them as input to randomly initialized bi-directional RNNs to capture the temporal dependencies across these frames. Libovický et al. [93] and Palaskar et al. [124] use ResNeXt-101 3D Convolutional Neural Network [51] trained to recognize 400 diverse human actions on the Kinetics dataset [74] to tackle the problem of generating text summaries for tutorial videos from How2 dataset [147].

Multi-modal fusion strategies: A lot of fusion techniques have been developed in the field of MMS. However, most of the works that take text-image-based inputs focus on *multi-modal attention* to facilitate a smooth information flow across the two modalities. Attention strategies have proven to be a very useful technique to help discard noise and focus on relevant information [173]. The

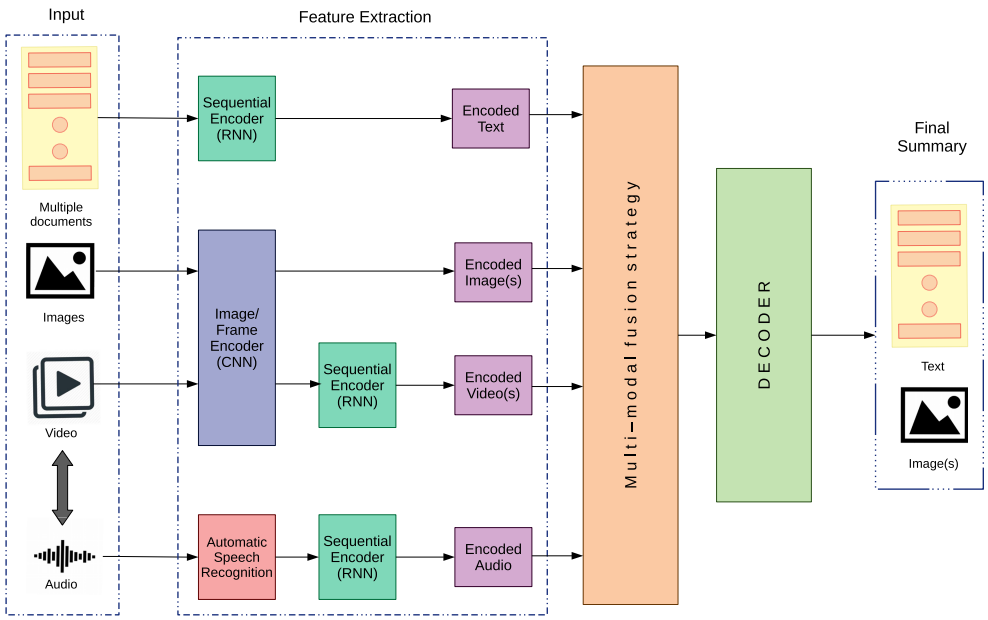


Fig. 4. A generic framework portraying existing neural models.

attention mechanism has been adopted by all the neural models that attempt to solve the MMS task. It has been applied to modal-specific information (*uni-modal attention*) as well as at the information-sharing step in the form of *multi-modal attention* to determine the degree of involvement of a specific modality for each input individually. Li et al. [87] proposed the hierarchical multi-modal attention for the first time to solve the task of multi-modal summarization of long sentences. The attention module comprises individual text and image attention layers, followed by a subsequent layer of modality attention layer. Although multi-modal attention has shown great promise in text-image summarization tasks, it itself is not sufficient for text-video-audio summarization tasks [92]. Hence, to overcome this weakness, Fu et al. [43] proposed *bi-hop attention* as an extension of bi-linear attention [79], and Li et al. [92] developed a novel *conditional self-attention mechanism* module to capture local semantic information of video conditioned on the input text information. Both of these techniques were backed empirically and established state-of-the-art in their respective problems.

Decoder: Depending on the encoding strategy used, the textual decoders also vary from plain *unidirectional RNN* [194] generating a word at a time to *hierarchical RNN decoders* [16] performing this step in multiple levels of granularity. Although a vast majority of neural models focus only on generating textual summary using multi-modal information as input [17, 86, 87, 93, 124], some work also output images as an supplement to the generated summary [16, 43, 92, 194, 195], reinforcing the textual information and improving the user experience. These works either use a post-processing strategy to select the image(s) to become a part of final multi-modal summary [16, 194] or they incorporate this functionality in their proposed model [43, 92, 195]. All the three frameworks that have implicit text-image summary generation characteristic adapt the final loss to be a weighted average of text generation loss together with the image selection loss. Zhu et al. [195] treat the image selection as a classification task and adopt cross-entropy loss to train the image selector. Fu et al. [43] also treat the image selection process as a classification problem and adopt an unsupervised learning technique that uses RL methods [192]. The proposed technique

uses representativeness and diversity as the two reward functions for the RL learning. Li et al. [92] propose a *cover frame selector*[13] that selects one image based on the hierarchical CNN-RNN-based video encoding conditioned on article semantics using a *conditional self-attention module*. Li et al. [92] use pairwise hinge loss to measure the loss during the model training.

Although the encoder-decoder model acts as the basic skeleton for the neural models solving MMS task, a lot of variations have been made, depending upon the input and output specifics. Zhu et al. [194] propose a visual coverage mechanism to mitigate the repetition of visual information. Li et al. [87] use two image filters, namely, *image attention filter* and *image context filter*, to avoid noise introduction, filtering out useful information. Zhu et al. [195] propose a multi-modal objective function that generates multi-modal summary at the end of this step, avoiding any statistical post-processing step for image selection. Fu et al. [43] utilize the fact that audio and video are synchronous, and audio can easily be converted to textual format, utilizing these speech transcriptions as the the bridge across the asynchronous modalities of text and video. They also formulate various fusion techniques including *early fusion* (concatenation of multi-modal embeddings), *tensor fusion* [185], and *late fusion* [99] to enhance the information representation in the latent space.

4.1.2 ILP-based Models. **Integer linear programming (ILP)** has been used for text summarization in the past [1, 44], primarily for extractive summarization. Jangra et al. [67] have shown that if properly formulated, then ILP can also be used to tackle the MMS task. More specifically, Jangra et al. [67] attempt to solve the problem of generating multi-modal summaries from a multi-document multi-modal news dataset by extracting necessary sentences, images, and videos. They propose a Joint Integer Linear Programming framework that optimizes weighted average of uni-modal salience and cross-modal correspondence. The model takes pre-trained joint embedding of sentences and images as input and performs a shared clustering, generating k_{txt} text clusters and k_{img} image clusters. A recommendation-based setting is used to create the most optimal clusters. The text cluster centers are chosen to be the extractive text summary, and a multi-modal summary containing text, images and videos is generated at the post-processing step.

4.1.3 Submodular Optimization-based Models. Sub-modular functions have been quite useful for text summarization tasks [95, 156], thanks to their assurance that the local optima is never worse than $1 - \frac{1}{e}$ (≈63%) of the global optima [119]. A greedy algorithm having time complexity of $O(nlogn)$ is sufficient to optimize the functions. Tiwari et al. [167], Li et al. [88], and Modani et al. [113] have also utilized these properties of submodular functions to solve the MMS task. Tiwari et al. [167] use *coverage*, *novelty*, and *significance* as the submodular functions to extract the most significant documents for the task of timeline generation of a social media event in a multi-modal setting. Li et al. [88] propose a linear combination of submodular functions (salience of text, redundancy and visual coverage, in this case) under a budget constraint to obtain near-optimal solutions at a sentence level to obtain an extractive text summary using news input comprising text, images, videos, and audio. Modani et al. [113] use a weighted sum of five submodular functions (coverage of input text/images, diversity of text/images in final summary, and coherence of text part and image part of the final summary) to generate a summary comprising text and images.

4.1.4 Nature-inspired Algorithms. Genetic algorithms [143] and other nature-inspired meta-heuristic optimization algorithms such as the Grey Wolf Optimizer [111] and Water Cycle algorithm [36] have shown great promise for extractive text summarization [144]. Jangra et al. [68] have illustrated that such algorithms can also be useful in multi-modal scenarios by experimenting with

[13]Model proposed by Li et al. [92] only selects one image per input, chosen from the video frames.

a multi-objective setting using differential evolution as the underlying guidance strategy. For the multi-objective optimization setup, the authors have proposed two different sets of objectives: one redundancy-based (including uni-modal salience, redundancy, and cross-modal correspondence) and one using cluster validity indices (PBM index [123] was used in this case). Both of these settings have performed better than the baselines. The optimization setup outputs the top most suitable sentences and images, which follow similar post-processing procedure as Jangra et al. [67]. Jangra et al. [69], however, used Grey Wolf Optimizer [111]-based multi-objective optimization strategy to obtain the combined complementary-supplementary multi-modal summaries. The proposed approach was split into two key steps: (a) **global coverage text format (GCTF)**, obtaining extractive text summaries using Grey Wolf Optimizer over all the input modalities in a clustering setup; (b) **visual enhanced text summaries (VETS)**, using one-shot population-based strategy to enhance the obtained text summaries with visual modalities to obtain the complementary and supplementary enhancements in a data-driven manner. The overall pipeline adopted similar pre-processing and post-processing steps as Jangra et al. [68].

4.1.5 Graph-based Models. Graph-based techniques have been widely adopted in extractive text summarization frameworks [33, 108, 109, 112]. These techniques involve graph formulation of text documents where nodes are represented by document sentences, and the edge weights are formulated using similarity across two sentences. Extending this idea to a multi-modal setup, Modani et al. [113] proposed a graph-based approach to generate text-image summaries. A graph was constructed using *content segments* (representing either sentences or images) as the nodes, and each node is given a weight depending on its information content. For sentences, this weight is computed as the sum of #nouns, #adverbs, #adjectives, #verbs, and half the #pronouns, while an image node's weight is given by the average similarity score with all other image segments. An edge weight for an edge connecting two sentences is computed as the cosine similarity of sentence embeddings (evaluated using auto-encoders), edge weight connecting two images is computed as the cosine similarity of image embeddings (evaluated using VGGNet [154]), and the edge weight connecting a sentence and an image is computed as the cosine similarity of sentence embedding and image embedding projected in a shared vector space (using Deep Fragment embeddings [72]). After graph construction, an iterative greedy strategy [112] is adopted to select appropriate content segments and generate the *text-image summary*.

Li et al. [88] also use a graph-based technique to evaluate the salience of text to generate an extractive text summary using multi-modal input (containing text documents, images, videos). A guided LexRank [33] was proposed to evaluate the salience score of the text unit (comprising document sentences and speech transcriptions). The guidance strategy proposed by Li et al. [88] had bidirectional connections for sentences belonging to documents, but only unidirectional connections were made for speech transcriptions with only outward edges to follow on their assumption that speech transcriptions might not always be grammatically correct, and hence should only be used for guidance and not for summary generation. This textual score was then used as a submodular function for the final model (refer to Section 4.1.3).

4.2 Post-processing

Most of the existing works are not capable of generating multi-modal summaries.[14] The systems that do generate multi-modal summaries either have an inbuilt system capable to generating multi-modal output (mainly by generating text using seq2seq mechanisms and selecting relevant

[14]Although all the surveyed methods are "multi-modal summarization" approaches, i.e., they all summarize multi-modal information, most of them summarize it to generate uni-modal outputs.

images) [92, 195] or they adopt some post-processing steps to obtain the visual and vocal supplements of the generated textual summaries [67, 194]. Neural network models that use multi-modal attention mechanisms to determine the relevance of modality for each input case have been used for selecting the most suitable image [16, 194]. More precisely, the visual coverage scores (after the last decoding step), i.e., the summation of attention values while generating the text summary, are used to determine the most relevant images. Depending upon the needs of the task, a single image [194] as well as multiple images [17] can be extracted to supplement the text.

Jangra et al. [67] propose a text-image-video summary generation task, which, as the name suggests, outputs all possible modalities in the final summary. Having extracted most important sentences and images (containing video key-frames as well) using the ILP framework, the images are separated from the key-frames and are supplemented with other images from the input set that have a moderate similarity with a pre-determined threshold and upper bound to avoid noisy and redundant information. Cosine similarity of global image features is used as the underlying similarity matrix in this case. A weighted average of verbal scores and visual scores is used to determine the most suitable video for the multi-modal summary. *verbal score* is defined as the information overlap between speech transcriptions and generated text summary, while the *visual score* is defined as the information overlap between the key-frames of a video with the generated image summary.

5 DATASETS AND EVALUATION TECHNIQUES

Due to the flexible nature of the MMS task, with a large variety of input-output modalities, the MMS task does not have a standard dataset used as a common evaluation benchmark for all approaches to this date. Nonetheless, we have collected information about datasets used in the previous works, and a comprehensive study of 20 datasets can be found at Table 3. It was found that out of these 21 datasets, 12 datasets are of news-related origin [43, 67, 87, 180, 194], and including the dataset on video tutorials by Sanabria et al. [147], there are 13 datasets that are domain-independent, thus suitable to test out domain-generic models. Six out of the 21 datasets produce text-only summaries using multi-modal input; out of these 6 datasets, 2 datasets' output comprises extracted text summaries [17, 88], and 4 datasets' output contains abstractive summaries [43, 86, 87, 147]. However, there are 8 datasets that output text-image summaries, which can further be divided into 6 extractive text-image summary generation datasets [10, 167, 178] and 2 abstractive text-image summary generation datasets [16, 92]. Datasets #19 [67] and #20 [69] are the only two datasets that comprise text, image, audio, and video in the output. However, these datasets are small and thus limited to extractive summarization techniques. Meanwhile, dataset #20 [69] is the only existing dataset that comprises both complementary and supplementary enhancements in the multi-modal summary (refer to Section 2 for the definition). Out of the 21 datasets, 17 datasets contain text in the multi-modal summary, 11 contain images as well, 3 comprise solely of audio-video outputs [38, 39, 146], and 1 dataset has a fixed template[15] as output [168]. Of these 17 text-containing datasets, 10 datasets contain extractive text summaries [10, 67, 167, 178] and the rest 7 datasets contain abstractive summaries [16, 43, 92, 147]. It is interesting to note that 5 out of these 7 abstractive datasets belong to the news-domain [16, 43, 87, 92, 194], while the other 2 focus on e-commerce product summarization [86] and tutorial summarization [147]. Out of the 21 datasets, only 4 [69, 88, 147, 194] of them are publicly available.

[15]Tjondronegoro et al. [168] focuses on summarizing tennis matches, and thus the output has a fixed template comprising three different summarization tasks: (a) summarization of entire tournament, (b) summarization of a match, and (c) summarization of a tennis player.

Depending on the input, we can also divide the 20 datasets based on the presence/absence of video in the input. There are 10 datasets that contain videos, whereas the rest 11 mostly work with text-image inputs. Due to the nature of this survey (the main focus on text modality), all 21 datasets in consideration contain text as input. A majority of these text sources are single documents [16, 86, 92, 194], but there are 6 datasets that have multiple documents in the input [10, 11, 67, 69, 88, 178]. Evangelopoulos et al. [39], Sanabria et al. [146], and Evangelopoulos et al. [38], however, do not contain text documents, but the speech transcriptions from corresponding audio inputs. While most of these datasets comprise multi-sentence summaries generated from input documents, the Li et al. [87] dataset contains a single sentence as the source as well as the reference summary. Most of these datasets use English-based text and audio, but there are 3 datasets that contain Chinese text [10, 11, 88, 92].

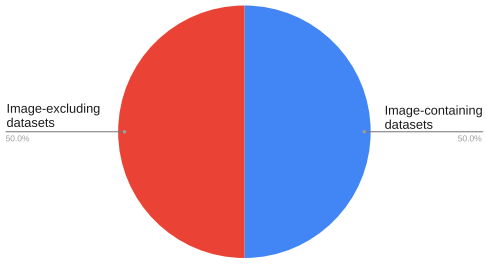
There are some datasets that have inputs other than text, image, audio, and video. For instance, Tiwari et al. [167] and Yan et al. [180] contain temporal information about the dataset for the task of multi-modal timelines generation. Qian et al. [132] also utilize user information including demographics such as gender, birthday, user profile (short biography), and other information, including user name, nickname, number of followers, number of microblogs posted, profile registration time, and user's levels of interest in different topics for generating summaries of an event based on social media content. Detailed plots for selected statistics on the datasets covered in this study can be found at Figure 5.

These datasets span a wide variety of domains, including sports such as tennis [168] and football [146], movies [38, 39], social media-based information [10, 11, 167], e-commerce [86]. In the coming days, we are likely bound to see more large-scale domain-specific datasets to advance this field.

Although there have been a lot of innovative attempts in solving the MMS task, the same does not go for the evaluation techniques used to showcase the quality of generated summaries. Most of the existing works use uni-modal evaluation metrics, including ROUGE scores [94] to evaluate the text summaries, accuracy, and precision-recall-based metrics to evaluate the image and video parts of generated summaries. A few works have also reported *True Positives* and *False Positives* as well [141]. The best way to evaluate the quality of a summary is to perform extensive human evaluations. Various techniques have been used to get the best user performance evaluations, including the quiz method [35] and user-satisfaction test [194]. These manual evaluation techniques are mainly of two kinds: (a) simple scoring of summary quality based on input [69, 194, 195] and (b) answering the questions based on input to quantify the information retention of input data instance [92]. However, one major issue with these manual evaluations is that they cannot be conducted for the entire dataset and are hence performed on a subset of the test dataset. There are a lot of uncertainties involving this subset, as well as the mental conditions of the human evaluators while performing these quality checks. Hence, it can be unreliable to compare two results of separate human evaluation experiments, even for the same task.

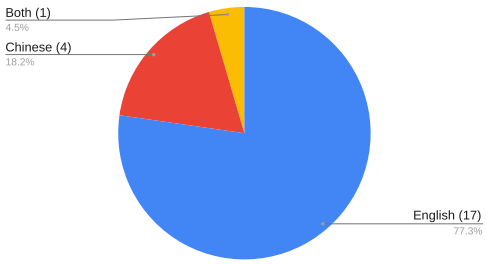
5.1 Text Summary Evaluation Techniques

Since the scope of this work is mostly limited to text-centric MMS techniques, it is important to discuss evaluation of text summaries separately and in tandem to other modalities. Even though quite a few MMS works generate uni-modal text summaries from multi-modal inputs [87, 88, 93], they still use very basic string-based n-gram overlap metrics like ROUGE [94] to conduct the evaluation. Through this survey, we want to direct the researchers to not just focus on ROUGE, but also look at other aspects of text summarization as well. For instance, Fabbri et al. [40] propose four key-characteristics that an ideal summary must have:

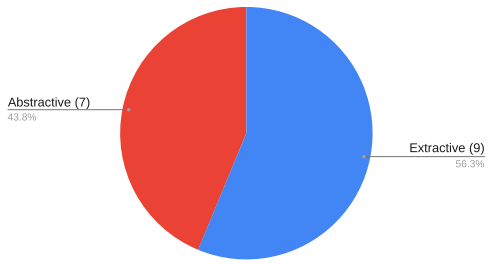


(a) Input image distribution

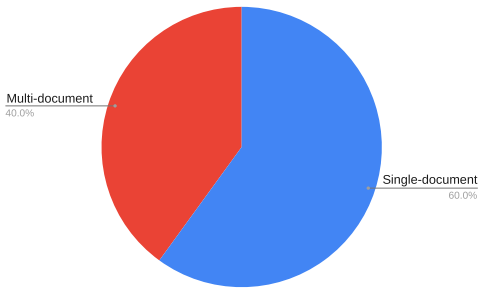
(b) Input audio/video distribution



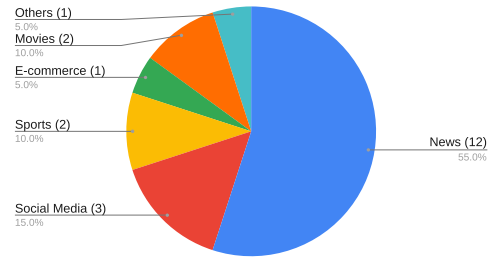
(c) Language distribution in datasets.



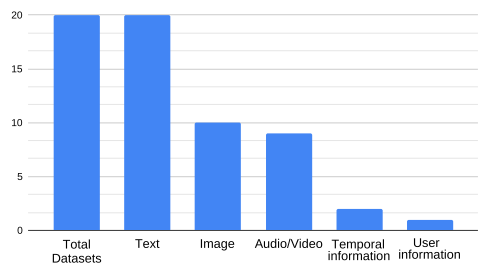
(d) Abstractive/Extractive text output



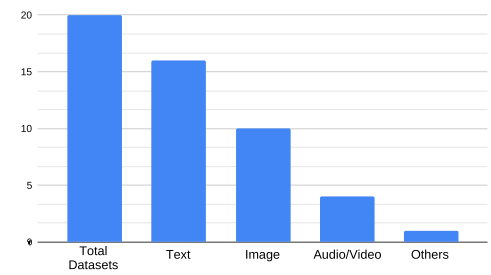
(e) Single vs Multiple document distribution for datasets



(f) Domain distribution for datasets



(g) Input modality distribution



(h) Output modality distribution

Fig. 5. Dataset statistics.

(1) **Coherence** the quality of smooth transition between different summary sentences such that sentences are not completely unrelated or completely same.
(2) **Consistency** the factual correctness of summary with respect to input document.
(3) **Fluency** the grammatical correctness and readability of sentences.
(4) **Relevance** the ability of a summary to capture important and relevant information from the input document.

Fabbri et al. [40] also illustrated how ROUGE is not capable of gauging the quality of generated summaries by doing an n-gram overlap with human-written reference summaries. There are other metrics out there that use more advanced strategies to do the evaluation, such as n-gram-based metric like WIDAR [65], embedding-based metrics such as ROUGE-WE [121], MoverScore [191], and **Sentence Mover Similarity (SMS)** [24] or neural model-based metrics such as BERTScore [190], SUPERT [46], BIANC [172], and S^3 [129]. These evaluation metrics have been proven to be empirically better at captivating the above-mentioned characteristics of a summary [40], and hence upcoming research works should also report performance on some of these metrics along with ROUGE to have more accurate analysis of the generated summaries.

5.2 Multi-modal Summary Evaluation Techniques

In an attempt to evaluate the multi-modal summaries, Zhu et al. [194] propose a **multi-modal automatic evaluation (MMAE)** technique that jointly considers uni-modal salience and cross-modal relevance. In their case, the final summary comprises text and images, and the final objective function is formulated as a mapping of three objective functions: (1) salience of text, (2) salience of images, and (3) text-image relevance. This mapping function is learned using supervised techniques (Linear Regression, Logistic Regression, and Multi-layer Perceptron, in their case) to minimize training loss with human judgment scores. Although the metric seems promising, there are a lot of conditions that must be met to perform the evaluation.

The MMAE metric does not effectively evaluate the information integrity[16] of a multi-modal summary, since it uses uni-modal salience scores as a feature in the overall judgment making process, leading to a cognitive bias. Zhu et al. [195] improve upon this by proposing an evaluation metric based on joint multi-modal representation (termed as MMAE++), projecting the generated summaries and the ground truth summaries into a joint semantic space. In contrast to other multi-modal evaluation metrics, they attempt to look at the multi-modal summaries as a whole entity, than a combination of piece-wise significant elements. A neural network-based model is used to train this joint representation. Images of two image caption pairs are swapped to obtain two image-text pairs that are semantically close to each other to obtain the training data for joint representation automatically. The evaluation model is trained using a multi-modal attention mechanism [87] to fuse the text and image vectors, using max-margin loss as loss function.

Modani et al. [113] propose a novel evaluation technique termed by them as ***Multimedia Summary Quality (MuSQ)***. Just like other multi-modal summarization metrics described above, *MuSQ* is also limited to text-image summaries. However, unlike the majority of previous evaluation metrics for multi-modal summarization or document summarization techniques [34], *MuSQ* does not require a ground truth to evaluate the quality of generated summary. *MuSQ* is a naive coverage-based evaluation metric denoted as μ_M and is defined as:

$$\mu_M = \mu_T + \mu_I + \sigma_{T,I} \tag{1}$$

[16]Information integrity is the dependability or trustworthiness of information. In context of a multi-modal summary evaluation task, it refers to the ability to make a judgment that is unbiased towards any modality (i.e., an ideal evaluation metric does not give higher importance to information from one modality (e.g., text) over other modality (e.g., images)).

Table 4. Comparative Study of Evaluation Techniques for Multi-modal Summarization

Metric name & corresponding paper	Pros & Cons
Multi-modal Automatic Evaluation (MMAE). Zhu et al. [194]	**Advantages** - MMAE shows high correlation with human judgment scores. **Disadvantages** - Requires a substantial manually annotated dataset. - Might perform ambiguously‡ for evaluation of other new domains.
MMAE++. Zhu et al. [195]	**Advantages** - Utilizes joint multi-modal representation of sentence-image pairs to better improve the correlation scores over MMAE metric [194]. **Disadvantages** - Requires a substantial manually annotated dataset. - Might perform ambiguously[1] for evaluation of other new domains.
Multimedia Summary Quality (MuSQ). Modani et al. [113]	**Advantages** - Does not require manually created gold summaries. **Disadvantages** - The technique is very naive and only considers coverage of input information and text-image cohesiveness. - The metric output is not normalized. Hence, the evaluation scores are highly sensitive to the cardinality of input text sentences and input images.

[1]Here, "might perform ambiguously" refers to the fact that, since model-based metrics are biased towards the training data, it is hard to determine how well they would perform on unseen domains. For instance, if the model is trained on news summarization dataset and the task is to evaluate medical report summaries, then the model performance cannot be determined without further experiments.

$$\mu_T = \sum_{v \in T} R_v * max_{u \in S}\{Sim(u, v)\} \tag{2}$$

$$\mu_I = \sum_{w \in V} \hat{R_w} * max_{x \in I}\{Sim(w, x)\} \tag{3}$$

$$\sigma_{T,I} = \sum_{v \in S} \sum_{w \in I} \{Sim(w, x) * R_v * \hat{R_w}\}, \tag{4}$$

where μ_T denotes the degree of coverage of input text document T by text summary S, μ_I denotes the degree of coverage of input images V by the image summary I. $\sigma_{T,I}$ measures the cohesion across the text sentences and images of final multi-modal summary. R_v and $\hat{R_w}$ are, respectively, the individual reward values for each input sentence and input image that denote the extent of information content in each content fragment (a text sentence or an image). A comparative study of evaluation techniques on Multi-modal Summarization can be found in Table 4.

To sum up, only a handful of works has focused on the evaluation of multi-modal summaries. Even the proposed evaluation metrics have a lot of drawbacks. The evaluation metrics proposed by Zhu et al. [194] and Zhu et al. [195] require a large human evaluation score-based training data to learn the parameter weights. Since these metrics are highly dependent on the human-annotated dataset, the quality of this dataset can compromise the evaluation process if the training dataset

is restrictive in domain coverage or is of poor quality. It also becomes difficult to generalize these metrics, since they depend on the domain of training data. The evaluation technique proposed by Modani et al. [113], although independent from gold summaries, is too naive and has its own drawbacks. The evaluation metric is not normalized and hence shows great variation when comparing the results of two input data instances with different sizes.

Overall, the discussed strategies have their own pros and cons; however, there is a great scope for future improvement in the area of "evaluation techniques for multi-modal summaries" (refer to Section 7).

6 RESULTS AND DISCUSSION

Since the MMS task is quite broad, covering multiple sub-problem statements, it is difficult to compare models due to the lack of a standard evaluation metric (refer to Section 5). We are then restricted to presenting the results using uni-modal evaluation techniques such as ROUGE scores [94] for text summaries and precision-recall scores for image summaries. In Section 3, we have described the diversity of works done so far, with some working on timeline generation [141, 167, 178], while others working on generic news summarization [68, 194], making it difficult to conduct a fair comparison of different architectures.[17] Even comparing two models that have very similar settings, such as Zhu et al. [194] and Chen and Zhuge [16] (both are trained on large-scale abstractive news summarization datasets), is not adequate, because datasets #5 and #7 have different sizes of training data (refer to Table 3). Other such examples are of Fu et al. [43] and Li et al. [92]; both these works intake text-video inputs; however, Fu et al. [43] is trained on English dataset with 2k instances, and Li et al. [92] is trained on Chinese dataset with 1.84k instances (refer to Table 3). Nonetheless, we attempted to give the readers an overview of the potential of existing architectures. There are a few observations that can be made even with these constraints. We can observe that the abstractive summarization models go neck-and-neck with extractive summarization models, even though extractive summarization models have an advantage of keeping the basic grammatical syntax intact, illustrating the advancement in neural summarization models in the MMS task. An extensive study can be found in Table 5.

There exist some works that share a common dataset to illustrate the efficacy of their proposed model architectures. For instance, Zhu et al. [194] and Zhu et al. [195] share a common dataset (dataset #2). Both the works produce competitive results, with Zhu et al. [195] outperforming Zhu et al. [194] by small difference in all modalities. It can also be observed from the results of Li et al. [88] that the input language does not affect the quality of summary at all. Results for both English and Chinese datasets (refer to dataset #3 in Table 3) are close, and the difference can be accredited to non-overlapping content across the two datasets. We can also observe from the results by Fu et al. [43] that neural models require large datasets to perform better. The CNN part of dataset only comprises 200 data instances, while the DailyMail part of dataset comprises 1,970 instances. The authors also suggest that the larger size of videos in CNN data leads to worse performance, even though the underlying learning strategies are the same.

Some datasets are also an extension of existing ones; for instance, dataset #19 [67] was extended from dataset #21 [88] by incorporating images and videos in the references, while dataset #20 [69] was extended from dataset #19 [67] by introducing complementary and supplementary enhancements for the multi-modal references. Therefore, all four works share the same reference summaries. Hence, even though the other modalities differ, the works can be partially compared with each other for the text modality. From this, it can be deduced that the two-step approach proposed by Reference [69] that first generates the ***Global Coverage Text Format summary (GCTF)***

[17]Note that we only display the results that have text as the *central modality* (refer to Section 2).

Table 5. Results of Different Methods for Text and Image Output Modalities

Paper	Dataset No.	Domain	Text score (ROUGE)				Image score			ME
			R-1	R-2	R-L	R-SU4	Precision	Recall	MAP	
Li et al. [88] (EXT)	Li et al. [88] (English)	News	0.442	0.133	N.A	0.187	N.A	N.A	N.A	✓
	Li et al. [88] (Chinese)		0.414	0.125	N.A	0.173	N.A	N.A	N.A	✓
Li et al. [87] (ABS)	Li et al. [87]	News	0.472	0.248	0.444	N.A	N.A	N.A	N.A	
Zhu et al. [194] (ABS)	Zhu et al. [194]	News	0.408	0.1827	0.377	N.A	0.624	N.A	N.A	✓
Zhu et al. [195] (ABS)	Zhu et al. [194]		0.411	0.183	0.378	N.A	0.654	N.A	N.A	✓
Chen and Zhuge [17] (EXT)	Chen and Zhuge [17]	News	0.271	0.125	0.156	N.A	N.A	N.A	N.A	
Chen and Zhuge [16] (ABS)	Chen and Zhuge [16]	News	0.326	0.120	0.238	N.A	N.A	0.4978	N.A	
Libovický et al. [93] (ABS)	Sanabria et al. [147]	Multi-domain	N.A	N.A	0.549	N.A	N.A	N.A	N.A	✓
Jangra et al. [67] (EXT)	Jangra et al. [67]	News	0.260	0.074	0.226	N.A	0.599	0.38	N.A	
Jangra et al. [68] (EXT)	Jangra et al. [67]		0.420	0.167	0.390	N.A	0.767	0.982	N.A	
Jangra et al. [69] (EXT)	Jangra et al. [69]	News	0.556	0.256	0.473	N.A	0.620	0.720	N.A	✓
Xu et al. [178] (EXT)	Xu et al. [178]	News	0.369	0.097	N.A	N.A	N.A	N.A	N.A	
Bian et al. [10] (EXT)	Bian et al. [10]	Social Media	0.507	0.303	N.A	0.232	N.A	N.A	N.A	
Yan et al. [180] (EXT)	Yan et al. [180]	News	0.442	0.109	0.320	N.A	N.A	N.A	N.A	
Bian et al. [11] (EXT)	Bian et al. [11] (social trends)	Social Media	0.504	0.307	N.A	0.235	N.A	N.A	N.A	
	Bian et al. [11] (product events)		0.478	0.279	N.A	0.187	N.A	N.A	N.A	
Fu et al. [43] (EXT)	Fu et al. [43] (DailyMail)	News	0.417	0.186	0.317	N.A	N.A	N.A	N.A	✓
	Fu et al. [43] (CNN)		0.278	0.088	0.187	N.A	N.A	N.A	N.A	✓
Li et al. [92] (ABS)	Li et al. [92]	News	0.251	0.096	0.232	N.A	N.A	N.A	0.654	✓
Li et al. [86] (ABS)	Li et al. [86] (Home Appliances)	E-commerce	0.344	0.125	0.224	N.A	N.A	N.A	N.A	✓
	Li et al. [86] (Clothing)		0.319	0.111	0.215	N.A	N.A	N.A	N.A	✓
	Li et al. [86] (Cases & Bags)		0.338	0.125	0.224	N.A	N.A	N.A	N.A	✓

This study is limited to works that contain text in the generated multi-modal summary.† Note that the comparison should be done with care, as most of the proposed approaches use different datasets (the "Dataset No." column corresponds to the ID column in Table 3). Column "ME" indicates presence/absence of manual evaluation in the corresponding work. Here, "N.A" or Not Available is used to denote the unavailablity of images in the output or unavailabilty of scores for an evaluation metric. "(ABS)" denotes abstractive summarization type, and "(EXT)" denotes extractive summarization type.

† For population-based techniques [68, 69], the best score across multiple solutions was reported in this work.

using Grey Wolf Optimizer on a multi-objective optimization setup, and then enhances this by using other modalities, outperforms all the prior works, illustrating the power of population-based techniques. The submodular optimization [88] is able to outperform the Genetic Algorithm technique [68], which is again a population-based-technique by some margin, which we believe can be credited to both the ability of sub-modular optimization as well as the tradeoff for the multi-modal summary generation framework. Since Li et al. [88] only generates text, while Jangra et al. [68] generates a multi-modal output comprising text, images, and videos; there might be some tradeoff to improve quality of other modalities over text. Since Jangra et al. [67] and Jangra et al. [68] both present their works on the same dataset (dataset #19), and it is evident that the population-based genetic algorithm proposed in Jangra et al. [68] produces better summaries as compared to the single point optimization strategy using integer linear programming proposed in Jangra et al. [67], both in terms of text as well as image output. For the video output[18] Jangra et al. [68] and Jangra et al. [67] performed equally well with an accuracy of 44%, while Jangra et al. [69] was able to obtain video accuracy of 64% (in contrast to the average accuracy of 16% for random selection over 10 attempts).

Out of the 17 works reported in Table 5, 8 have performed some sort of manual evaluation along with automatic evaluation to produce a clearer picture about the performance of various summarization strategies. Through these experiments, prior works have statistically shown how the presence of multi-modal information can not only aid the uni-modal summarization process, but improve the overall user experience. Li et al. [92] have shown that an output containing text and

[18]Since dataset #19 [67] and #20 [69] are the only datasets that contain text and video in the output, we have reported the results in text instead of making another column in Table 5. It should also be noted that accuracy is used to evaluate the video summary, because both of these datasets restrict a single video in the output summary. Since dataset #20 is extended from dataset #19, they both share the same text and video outputs.

images increases user satisfaction by 12.4% in juxtaposition to text summaries. Jangra et al. [69] also illustrate that having visual cues in a text summary helps improve the overall satisfaction by 22%, makes the topic 19% more fascinating, and helps users understand the topic better by 14.5%. Jangra et al. [69] also empirically justify through manual annotations that a multi-modal summary should have both complementary and supplementary enhancements to improve the user experience.

7 FUTURE WORK

The MMS task is relatively new, and the work done so far has only scratched the surface of what this field has to offer. In this section, we discuss the future scope of the MMS task, including some possible improvements in existing works, as well as some possible new directions.

7.1 Scope of Improvement

Better fusion of multi-modal information: Almost all the works discussed in this survey adopt a late-joint representation approach, where uni-modal information is extracted beforehand, and the information-sharing across multiple modalities takes place at a later stage. These works either use a pre-trained model on image captions [154] or train the multi-modal correspondence in a naïve way, using a neural multi-modal attention mechanism. However, Liu et al. [100] have proposed a multi-stage fusion approach with a fusion forget gate module for solving the task of multimodal summarization in videos. Their proposed approach tries to improve the interaction between multiple modalities to complete the missing information of each modality. Further, they have also introduced a forget gate to suppress the flow of unecessary multimodal noise. Using this approach, the model was able to outperform the Palaskar et al. [124] model 8.3 BLEU-4 points, 7.4 ROUGE points, and 3.9 METEOR points in the How2 [147] dataset. Although these techniques are able to capture the essence of semantic overlap across modalities, there is still room for improvement in fusion modeling.

Better evaluation metrics (for multi-modal summaries): Most of the existing works use uni-modal evaluation techniques such as ROUGE scores [94] for text and precision-recall-based metrics for images and videos. The multi-modal evaluation metrics proposed by Zhu et al. [194] and Zhu et al. [195] have shown some promise, but they require a large set of human evaluation scores of generated summaries for training to determine the parameter values, making them unfit as a universal metric, especially when the summaries to be evaluated are from different domains than the data the models were trained on. These proposed evaluation metrics are also very specific, as they work only for text-image-based multi-modal summaries. Hence, the community still lacks an evaluation metric that could judge the quality of a summary comprising multiple summaries. Even the standard text summarization metrics have some inherent shortcomings, as illustrated by the survey performed by Ter Hoeve et al. [166]. They illustrated that even though these metrics are able to cover up basic concepts such as informativeness, fluency, succinctness, and factuality, they still lack other important aspects like usefulness as discovered by the survey conducted on users who frequently use automatic summarization mechanisms. To improve the overall user satisfaction, similar techniques should be incorporated for the evaluation of multi-modal summarization systems as well.

More datasets: All the datasets proposed in the community to date are mostly centered towards the news domain, even though there are multiple potential applications in other domains, such as medical report summarization, tutorial summarization, simplification summarization, slogan generation, and so on, that could benefit from multi-modal information. There are also potential new research areas that can be explored, but due to the lack of dataset availability, the community

is unable to pursue research in these fields. Some of these are: explainable MMS, sentiment lossless MMS, multi-lingual MMS, data-stream MMS, large-scale MMS of long documents and so on.

Complementary and Supplementary MMS: It is well-established fact that multi-modal systems improve the user experience and help paint a clearer picture of topics or events discussed in input documents [69, 92]. However, there does not exist any system that can generate the complementary and supplementary multi-modal summaries together. A large majority of research work today focuses on developing supplementary multi-modal summaries [16, 194]. There also exist some works that generate complementary multi-modal summaries as well [92]. Jangra et al. [69] also illustrated how an ideal multi-modal summary should comprise both complementary and supplementary enhancements.

But the concepts of complementary and supplementary enhancements should not be limited to visual modalities over textual central modality as proposed in Jangra et al. [69]. For instance, summarizing articles with user opinion from the comments section can be a great application.[19] Even though this is a text-only task, the concepts of complementary and supplementary enhancements can be extended to cover up comments that cover vivid perspectives, in both favor and against the information presented in the article.

No abstractive complementary-supplementary MMS framework or application has been proposed in the community so far, and hence the exploration potential in this is quite vast.

7.2 New Directions

Manually generated dataset for evaluation of MMS evaluation metrics: There is a need of some human-annotated datasets to evaluate the performance of existing and upcoming evaluation metrics. There have been some works in text summarization that can be used to draw out some parallels; for instance, Fabbri et al. [40] released the SummEval dataset that gives out human-annotation scores for 1,600 article documents scored by 11 annotators in four key-characteristics of a summary: *consistency*, *coherence*, *fluency,* and *relevance*. Similar work is also needed in the MMS, where, other than uni-modal aspects, the ability to judge the cross-modal information correspondence also should be taken into account.

Explainable and Controlled MMS: Maynez et al. [106] showed that automated abstractive summarization models suffer from the problem of hallucinations and often generate fictional content. Explainable and Controlled MMS refers to the process of developing summarization systems where we do not treat these automated systems as black boxes generating summaries; rather, we have the power to understand and control the output of these models to produce content of our desired type. Even though existing MMS summarization frameworks have shown substantial improvement in the recent few years, it is still a mystery how each modality is handled and understood to obtain the final summaries. This calls for more explainable systems that also output some meta-data in tandem with the summaries to better understand the functioning of these models. Attention mechanism [5] is one way to get better insights in the model working. In the context of text summarization, Haonan et al. [50] proposed a select-and-generate strategy where elements are first extracted from a document based on informativeness, novelty, and relevance and then an abstractor generates an abstractive summary using the extracted elements. Their extractor module features an interaction matrix to explain the selection logic, and by changing the thresholds of the model, one can control the final summary quality.

[19]This task can be considered multi-modal if we extend the notion of modality to something more generic; however, since the scope of this survey is limited to the distinction in modality being the form of representing information, we do not consider such works in great detail.

In the multimodal context, Shang et al. [152] proposed DGExplain, which exploits the cross-modal association between the news of multiple modalities and the user comments to detect misinformation. Explainable and controlled multimodal summarization systems can be built using this kind of explainable framework to detect and filter incorrect content and summarize the true facts. Mukherjee et al. [118] proposed a multi-tasking approach to generate topic-aware multimodal summaries. Their proposed model aims to embed topic awareness in both the visual and textual outputs. Thus, these kinds of models are stepping stones towards developing systems that are able to control the information flow from different modalities in input and output.

Application-oriented MMS: We can use different MMS techniques to leverage the output of various tasks, such as product description generation, product review summarization, multi-modal microblog summarization, education material summarization, medical report summarization, and simplification of any multi-modal content. For each of these tasks, earlier text-only [2, 20, 184] or image-only [159] summarization methods were majorly used. However, Li et al. [86] showed that the quality of e-commerce product descriptions could be improved by incorporating visual information and textual descriptions of a product during the summarization process. Delbrouck et al. [27] utilized the visual features from the x-rays associated with the radiology reports to improve the medical report summarization quality.

During any natural disaster, people post relevant content on microblogging websites, which concerned authorities could use for rescue operations. Saini et al. [142] proposed a multi-modal approach to summarize these posts utilizing both the textual and visual aspects of the post to improve the summary quality. Recently, educational content has been multi-modal, comprising video, audio, and text. We believe that educational material summary quality can be significantly enhanced if information from all of these modalities is utilized during the summarization process [77]. All of these recent works highlight the ability of MMS to combine the knowledge from various modalities to produce superior-quality summaries; hence, making it a more robust choice over the traditional uni-modality-based methods for multiple applications in future.

Sentiment/Emotion Lossless MMS: The point of a summary is to provide users with the information that they would gain from reading the entire document; and an ideal summary would not only do that, but also elicit the same sentiments that the user would feel when reading the entire document. Gulshan et al. [47] propose an extractive text summarization framework that attempts to retain the sentiment of input in the generated summary. This task would be very relevant in a few domains, such as story summarization, novel summarization, and so on, where the users tend to empathize with the content in the summary. Khan and Fu [76] proposed a transformer-based architecture to perform aspect-based multimodal sentiment analysis. In the future, we can combine ideas from aspect-based summarization systems [86] and multimodal aspect-based sentiment recognition frameworks to generate sentiment-aware MMS. When working with multi-modal data, this becomes even more challenging and interesting, since various additional flavors of sentiment can be obtained from different modalities; and in some cases, some modalities can fill the lack of sentiment in others. For instance, in a news article covering an earthquake, the text tends to be objective and devoid of subjective and sentiment-bearing expressions to remain professional, but the images and videos are able to convey these sentiments and emotions conveniently. Hence, we believe that this kind of multi-modal summarization would help move current systems one step further in an attempt to obtain ideal summaries.[20]

[20]Note that this problem would be mostly restrictive to single-document summarization tasks (with some exceptions), since multiple articles tend to cover different aspects of a topic, often leading to conflicting opinions, and hence conflicting sentiments and emotions. There, the problem statement can be changed to providing an unbiased and sentiment-less summary to be faithful to the users.

Multi-lingual MMS: Multi-modal information has proven to be useful for multi-modal neural machine translation tasks [132, 160], and it has been a highly debated question whether language affects visual perception, a universal form of perception shared by all individuals [175]. The fact that this question remains open to this date speaks volumes about how multi-modal information can prove to be useful for multi-lingual summarization tasks, if harnessed properly.

Data-stream MMS: Data-stream summarization, also known as *update summarization* or *online summarization* or *dynamic summarization,* has been explored in great extent in the automatic text summarization community [48, 57, 98, 140, 153, 164, 170, 187]. Data-stream summarization is used in situations where the input information is not static and, accordingly, the summarization system needs to dynamically keep the summary up-to-date with the latest information. It is a challenging problem, as it requires the summary to retain the key-highlights from past events while being consistent and fluent with the most recent events as well. Data-stream summarization has been used for various applications, such as social media content summarization [98, 153], review summarization [48, 57, 170, 187], and so on.

With the world moving towards multi-modal information representation, there is a need to make these models robust and adaptive to multi-modal information. A few of these are discussed in the "Application-oriented MMS" part of this section.

Query-based MMS: A lot of work has been done in query-based text summarization [97, 133], but there is no existing research on query-based summarization in a multi-modal setting. Since it has been shown that visual content can help improve the quality of experience [194], we believe that query-based summarization setup, which has a user interaction, could really be improved by introducing multi-modal form of information.

MMS at scale: Although some work has been done on generic datasets in terms of domain coverage [16, 68, 88, 194], most of the existing works have been performed in a protective environment with some pre-defined notions of input and output formats. To produce a large-scale ready-to-use MMS framework, a more generic setup is required that has better generalization and high adaptive capabilities.

MMS with user interaction: Inspired from query-chain summarization frameworks [9], there is a possibility of a multi-modal summarization based on user interaction, which could help improve the overall user satisfaction.

8 CONCLUSION

Due to the improving technology, it has become convenient for people to create and share information in multiple modalities, a feat that was not possible a decade ago. As a result of this advancement, the need for multi-modal summarization is increasing. We present a survey to help familiarize users with techniques and challenges present for the MMS task. In this manuscript, we formally define the task of multi-modal summarization, and we also provide an extensive categorization of existing works depending upon various input-, output-, and technique-related details. We then include a comprehensive description of datasets used to tackle the MMS task. Moreover, we also briefly describe various techniques used to solve the MMS task, along with the evaluation metrics used to judge the quality of summaries produced. Finally, we also provide a few possible directions that research in MMS can take. We hope that this survey article will significantly promote research in multi-modal summarization.

REFERENCES

[1] Rasim Alguliev, Ramiz Aliguliyev, and Makrufa Hajirahimova. 2010. Multi-document summarization model based on integer linear programming. *Intell. Contr. Autom.* 1, 2 (2010), 105.

[2] Syed Muhammad Ali, Zeinab Noorian, Ebrahim Bagheri, Chen Ding, and Feras Al-Obeidat. 2020. Topic and sentiment aware microblog summarization for Twitter. *J. Intell. Inf. Syst.* 54, 1 (2020), 129–156.
[3] Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 337–342.
[4] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multim. Syst.* 16, 6 (2010), 345–379.
[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:cs.CL/1409.0473.
[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2018), 423–443.
[7] Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. 2018. Multimodal emoji prediction. *arXiv preprint arXiv:1803.02392*.
[8] Madhushree Basavarajaiah and Priyanka Sharma. 2019. Survey of compressed domain video summarization techniques. *ACM Comput. Surv.* 52, 6 (2019), 1–29.
[9] Tal Baumel, Raphael Cohen, and Michael Elhadad. 2014. Query-chain focused summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 913–922.
[10] Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia summarization for trending topics in microblogs. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. 1807–1812.
[11] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2014. Multimedia summarization for social events in microblog stream. *IEEE Trans. Multim.* 17, 2 (2014), 216–228.
[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, Jan. (2003), 993–1022.
[13] Teresa Botschen, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly-Sergieh, and Stefan Roth. 2018. Multimodal frame identification with multilingual evaluation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1481–1491.
[14] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.
[15] Fan Chen, Christophe De Vleeschouwer, H. Duxans Barrobés, J. Gregorio Escalada, and David Conejero. 2010. Automatic summarization of audio-visual soccer feeds. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 837–842.
[16] Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 4046–4056.
[17] Jingqiang Chen and Hai Zhuge. 2018. Extractive text-image summarization using multi-modal RNN. In *Proceedings of the 14th International Conference on Semantics, Knowledge and Grids (SKG)*. IEEE, 245–248.
[18] Jingqiang Chen and Hai Zhuge. 2019. News image captioning based on text summarization using image as query. In *Proceedings of the 15th International Conference on Semantics, Knowledge and Grids (SKG)*. IEEE, 123–126.
[19] Jingqiang Chen and Hai Zhuge. 2022. A news image captioning approach based on multimodal pointer-generator network. *Concurrency and Computation: Practice and Experience* 34, 7 (2022), e5721.
[20] Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3040–3050.
[21] Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 675–686.
[22] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. arXiv:cs.CV/1909.11740.
[23] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.
[24] Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P19-1264
[25] Andrei Catalin Coman, Yaroslav Nechaev, and Giacomo Zara. 2018. Predicting emoji exploiting multimodal data: FBK participation in ITAmoji task. *EVALITA Eval. NLP Speech Tools Ital.* 12 (2018), 135.
[26] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. 2019. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*. IEEE, 2090–2096.

[27] Jean-Benoit Delbrouck, Cassie Zhang, and Daniel Rubin. 2021. QIAI at MEDIQA 2021: Multimodal radiology report summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. 285–290.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

[31] Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2974–2978.

[32] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177* (2017).

[33] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *J. of Artif. Intell. Res.* 22 (2004), 457–479.

[34] Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Inf. Process. Manag.* 56, 5 (2019), 1794–1814.

[35] Berna Erol, D.-S. Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *Proceedings of the International Conference on Multimedia and Expo*. IEEE, III–25.

[36] Hadi Eskandar, Ali Sadollah, Ardeshir Bahreininejad, and Mohd Hamdi. 2012. Water cycle algorithm–A novel metaheuristic optimization method for solving constrained engineering optimization problems. *Comput. Struct.* 110 (2012), 151–166.

[37] Georgios Evangelopoulos, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, A. Zlatintsi, and Yannis Avrithis. 2008. Movie summarization based on audiovisual saliency detection. In *Proceedings of the 15th IEEE International Conference on Image Processing*. IEEE, 2528–2531.

[38] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimed.* 15, 7 (2013), 1553–1568.

[39] Georgios Evangelopoulos, Athanasia Zlatintsi, Georgios Skoumas, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, and Yannis Avrithis. 2009. Video event detection and summarization using audio, visual and text saliency. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3553–3556.

[40] A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, R. Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Trans. Assoc. Computat. Ling.* 9 (2021), 391–409.

[41] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. arXiv:cs.CL/2007.01852.

[42] Julian Fierrez-Aguilar, Javier Ortega-Garcia, Joaquin Gonzalez-Rodriguez, and Josef Bigun. 2005. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recog.* 38, 5 (2005), 777–779.

[43] Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multi-modal summarization for video-containing documents. *arXiv preprint arXiv:2009.08018*.

[44] Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*. 911–926.

[45] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: A survey. *Artif. Intell. Rev.* 47, 1 (2017), 1–66.

[46] Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.acl-main.124

[47] Varun Gulshan, Renu P. Rajan, Kasumi Widner, Derek Wu, Peter Wubbels, Tyler Rhodes, Kira Whitehouse, Marc Coram, Greg Corrado, Kim Ramasamy, Rajiv Raman, Lily Peng, and Dale R. Webster. 2019. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalm.* 137, 9 (09 2019), 987–993. DOI: https://doi.org/10.1001/jamaophthalmol.2019.2004

[48] Pankaj Gupta, Ritu Tiwari, and Nirmal Robert. 2016. Sentiment analysis and text summarization of online reviews: A survey. In *Proceedings of the International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 0241–0245.

[49] Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *J. Emerg. Technol. Web Intell.* 2, 3 (2010), 258–268.
[50] Wang Haonan, Gao Yang, Bai Yu, Mirella Lapata, and Huang Heyan. 2020. Exploring explainable selection to control abstractive summarization. *arXiv preprint arXiv:2004.11779*.
[51] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6546–6555.
[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[53] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computat.* 9, 8 (1997), 1735–1780.
[54] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* 47 (2013), 853–899.
[55] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision*. 4193–4202.
[56] Chiori Hori, Takaaki Hori, Gordon Wichern, Jue Wang, Teng-yok Lee, Anoop Cherian, and Tim K. Marks. 2018. Multimodal attention for fusion of audio and spatiotemporal features for video description. In *Proceedings of the CVPR Workshops*. 2528–2531.
[57] Ya-Han Hu, Yen-Liang Chen, and Hui-Ling Chou. 2017. Opinion mining from online hotel reviews–a text summarization approach. *Inf. Process. Manag.* 53, 2 (2017), 436–449.
[58] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the 1st Conference on Machine Translation*. 639–645.
[59] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 172–189.
[60] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.
[61] Tanveer Hussain, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C. de Albuquerque. 2020. A comprehensive survey of multi-view video summarization. *Pattern Recog.* 109 (2020), 107567.
[62] M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain. 2003. Multimodal biometric authentication methods: A COTS approach. In *Proceedings of the Workshop on Multimodal User Authentication*. Citeseer, 99–106.
[63] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* 108, 1-2 (2007), 116–134.
[64] Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization. *arXiv preprint arXiv:2212.01669* (2022).
[65] Raghav Jain, Vaibhav Mavi, Anubhav Jangra, and Sriparna Saha. 2022. WIDAR—Weighted input document augmented ROUGE. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Springer-Verlag, Berlin, 304–321. DOI: https://doi.org/10.1007/978-3-030-99736-6_21
[66] Anubhav Jangra, Raghav Jain, Vaibhav Mavi, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Semantic extractor-paraphraser based abstractive summarization. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. 191–199.
[67] Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. 2020. Text-image-video summary generation using joint integer linear programming. In *Proceedings of the European Conference on Information Retrieval*. Springer, 190–198.
[68] Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020. Multi-modal summary generation using multi-objective optimization *(SIGIR'20)*. Association for Computing Machinery, New York, NY, 1745–1748. DOI: https://doi.org/10.1145/3397271.3401232
[69] Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammed Hasanuzzaman. 2021. *Multi-modal Supplementary-Complementary Summarization Using Multi-objective Optimization*. Association for Computing Machinery, New York, NY, 818–828. DOI: https://doi.org/10.1145/3404835.3462877
[70] Hira Javed, M. M. Sufyan Beg, and Nadeem Akhtar. 2022. Multimodal summarization: A concise review. In *Proceedings of the International Conference on Computational Intelligence and Sustainable Technologies*. Springer, 613–623.
[71] Prince Jha, Gaël Dias, Alexis Lechervy, Jose G. Moreno, Anubhav Jangra, Sebastião Pais, and Sriparna Saha. 2022. Combining vision and language representations for patch-based identification of lexico-semantic relations. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4406–4415.
[72] Andrej Karpathy, Armand Joulin, and Li F. Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1889–1897.

[73] Tsuneaki Kato. 2021. Multi-modal summarization. In *Evaluating Information Retrieval and Access Tasks*. Springer, Singapore, 71–82.
[74] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
[75] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*.
[76] Zaid Khan and Yun Raymond Fu. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia*.
[77] Aman Khullar and Udit Arora. 2020. MAST: Multimodal abstractive summarization with trimodal hierarchical attention. *arXiv preprint arXiv:2010.08021*.
[78] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual QA. *Adv. Neural Inf. Process. Syst.* 29 (2016), 361–369.
[79] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.
[80] Sujeong Kim, David Salter, Luke DeLuccia, Kilho Son, Mohamed R. Amer, and Amir Tamrakar. 2018. SMILEE: Symmetric multi-modal interactions with language-gesture enabled (AI) embodiment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. 86–90.
[81] Mahesh Kini and Karthik Pai. 2019. A survey on video summarization techniques. In *Proceedings of the Conference on Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, 1–5.
[82] Elsa Andrea Kirchner, Marc Tabie, and Anett Seeland. 2014. Multimodal movement prediction-towards an individual assistance of patients. *PloS One* 9, 1 (2014), e85060.
[83] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid Gaussian-Laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*.
[84] Theodoros Kostoulas, Guillaume Chanel, Michal Muszynski, Patrizia Lombardo, and Thierry Pun. 2017. Films, affective computing and aesthetic experience: Identifying emotional and aesthetic highlights from multimodal signals in a social setting. *Front. ICT* 4 (2017), 11.
[85] Yaniv Leviathan and Yossi Matias. 2018. Google Duplex: An AI system for accomplishing real-world tasks over the phone.
[86] Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Aspect-aware multimodal summarization for Chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8188–8195.
[87] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 4152–4158. DOI: https://doi.org/10.24963/ijcai.2018/577
[88] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1092–1102.
[89] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Trans. Knowl. Data Eng.* 31, 5 (2018), 996–1009.
[90] Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. MAEC: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3063–3070.
[91] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
[92] Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. VMSMO: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406*.
[93] Jindrich Libovickỳ, Shruti Palaskar, Spandana Gella, and Florian Metze. 2018. Multimodal abstractive summarization for open-domain videos. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*.
[94] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. Retrieved from https://www.aclweb.org/anthology/W04-1013.
[95] Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 912–920.

[96] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. 2021. VX2TEXT: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7005–7015.

[97] Marina Litvak and Natalia Vanetik. 2017. Query-based summarization using MDL principle. In *Proceedings of the MultiLing Workshop on Summarization and Summary Evaluation across Source Types and Genres*. 22–31.

[98] Cheng-Ying Liu, Ming-Syan Chen, and Chi-Yao Tseng. 2015. IncreSTS: Towards real-time incremental short text summarization on comment streams from social network services. *IEEE Trans. Knowl. Data Eng.* 27, 11 (2015), 2986–3000.

[99] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*.

[100] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1834–1845. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.144

[101] Xin Liu and Yujia Jiang. 2021. Aesthetic assessment of website design based on multimodal fusion. *Fut. Gen. Comput. Syst.* 117 (2021), 433–438.

[102] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2890–2903.

[103] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 13–23.

[104] Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Devel.* 2, 2 (1958), 159–165.

[105] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2020. Multi-document Summarization via Deep Learning Techniques: A Survey. arXiv:cs.CL/2011.04843.

[106] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

[107] Lianhai Miao, Da Cao, Juntao Li, and Weili Guan. 2020. Multi-modal product title compression. *Inf. Process. Manag.ement* 57, 1 (2020), 102123.

[108] Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 170–173.

[109] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 404–411.

[110] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 3111–3119.

[111] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. 2014. Grey Wolf Optimizer. *Adv. Eng. Softw.* 69 (2014), 46–61.

[112] Natwar Modani, Elham Khabiri, Harini Srinivasan, and James Caverlee. 2015. Creating diverse product review summaries: A graph approach. In *Proceedings of the International Conference on Web Information Systems Engineering*. Springer, 169–184.

[113] Natwar Modani, Pranav Maneriker, Gaurush Hiranandani, Atanu R. Sinha, Vaishnavi Subramanian, Shivani Gupta, et al. 2016. Summarizing multimedia content. In *Proceedings of the International Conference on Web Information Systems Engineering*. Springer, 340–348.

[114] Arthur G. Money and Harry Agius. 2008. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Commun. Image Represent.* 19, 2 (2008), 121–143.

[115] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2000–2008.

[116] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*.

[117] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. 169–176.

[118] Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. Topic-aware multimodal summarization. In *Findings of the Association for Computational Linguistics (AACL-IJCNLP'22)*. 387–398.

[119] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions–I. *Math. Program.* 14, 1 (1978), 265–294.

[120] Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining Text Data*. Springer, 43–76.
[121] Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1925–1930. DOI: https://doi.org/10.18653/v1/D15-1222
[122] Payam Oskouie, Sara Alipour, and Amir-Masoud Eftekhari-Moghadam. 2014. Multimodal feature extraction and fusion for semantic mining of soccer video: A survey. *Artif. Intell. Rev.* 42, 2 (2014), 173–210.
[123] Malay K. Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. 2004. Validity index for crisp and fuzzy clusters. *Pattern Recog.* 37, 3 (2004), 487–501.
[124] Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901*.
[125] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4055–4064.
[126] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*.
[127] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *Lond., Edinb. Dubl. Philos. Mag. J. Sci.* 2, 11 (1901), 559–572.
[128] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.
[129] Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W17-4510
[130] Xueming Qian, Mingdi Li, Yayun Ren, and Shuhui Jiang. 2019. Social media based event summarization by user–text–image co-clustering. *Knowl.-based Syst.* 164 (2019), 107–121.
[131] Xueming Qian, Yao Xue, Xiyu Yang, Yuan Yan Tang, Xingsong Hou, and Tao Mei. 2014. Landmark summarization with diverse viewpoints. *IEEE Trans. Circ. Syst. Vid. Technol.* 25, 11 (2014), 1857–1869.
[132] Xin Qian, Ziyi Zhong, and Jieli Zhou. 2018. Multimodal machine translation with reinforcement learning. *arXiv preprint arXiv:1805.02356*.
[133] Nazreena Rahman and Bhogeswar Borah. 2020. Improvement of query-based text summarization using word sense disambiguation. *Complex Intell. Syst.* 6 (2020), 75–85.
[134] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Sig. Process. Mag.* 34, 6 (2017), 96–108.
[135] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *Proceedings of the 24th ACM International Conference on Multimedia*. 1092–1096.
[136] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:cs.CV/2102.12092.
[137] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 139–147.
[138] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of Spanish online videos. *IEEE Intell. Syst.* 28, 3 (2013), 38–45.
[139] Nils Ryden, Choon B. Park, Peter Ulriksen, and Richard D. Miller. 2004. Multimodal approach to seismic pavement testing. *J. Geotechnic. Geoenviron. Eng.* 130, 6 (2004), 636–645.
[140] Horacio Saggion and Thierry Poibeau. 2013. Automatic text summarization: Past, present and future. In *Multi-source, Multilingual Information Extraction and Summarization*. Springer, 3–21.
[141] Mathilde Sahuguet and Benoit Huet. 2013. Socially motivated multimedia topic timeline summarization. In *Proceedings of the 2nd International Workshop on Socially-aware Multimedia*. 19–24.
[142] Naveen Saini, Sriparna Saha, Pushpak Bhattacharyya, Shubhankar Mrinal, and Santosh Kumar Mishra. 2021. On multimodal microblog summarization. *IEEE Trans. Computat. Soc. Syst.ems* 9, 5 (2021), 1317–1329.
[143] Naveen Saini, Sriparna Saha, Dhiraj Chakraborty, and Pushpak Bhattacharyya. 2019. Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures. *PloS One* 14, 11 (2019), e0223477.
[144] Naveen Saini, Sriparna Saha, Anubhav Jangra, and Pushpak Bhattacharyya. 2019. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, Grey Wolf Optimizer and water cycle algorithm. *Knowl.-based Syst.* 164 (2019), 45–67.

[145] Gerard Salton. 1989. Automatic text processing: The transformation, analysis, and retrieval of. 169 (1989). https://dl.acm.org/doi/book/10.5555/77013.
[146] Melissa Sanabria, Frédéric Precioso, and Thomas Menguy. 2019. A deep architecture for multimodal summarization of soccer games. In *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*. 16–24.
[147] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.
[148] Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia*. 456–465.
[149] Tinumol Sebastian and Jiby J. Puthiyidam. 2015. A survey on video summarization techniques. *Int. J. Comput. Appl* 132, 13 (2015), 30–32.
[150] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S. Huang. 2005. Multimodal approaches for emotion recognition: A survey. In *Internet Imaging VI*, Vol. 5670. International Society for Optics and Photonics, 56–67.
[151] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR* abs/1704.04368.
[152] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A duo-generative approach to explainable multimodal COVID-19 misinformation detection. In *Proceedings of the ACM Web Conference*. 3623–3631.
[153] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: Continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 533–542.
[154] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.
[155] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5317–5332.
[156] Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 224–233.
[157] Robert Snelick, Umut Uludag, Alan Mink, Mike Indovina, and Anil Jain. 2005. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 3 (2005), 450–455.
[158] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image Vis Comput.* 65 (2017), 3–14.
[159] S. K. Somasundaram and P. Alli. 2017. A machine learning ensemble classifier for early prediction of diabetic retinopathy. *J. Med. Syst.* 41, 12 (2017), 1–12.
[160] Lucia Specia. 2018. Multi-modal context modelling for machine translation. (2018). https://rua.ua.es/dspace/handle/10045/76101.
[161] Chanchal Suman, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Why pay more? A simple and efficient named entity recognition system for tweets. *Exp. Syst. Applic.* (2020), 114101.
[162] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 7464–7473.
[163] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.
[164] Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. 2011. Summarizing a document stream. In *Proceedings of the European Conference on Information Retrieval*. Springer, 177–188.
[165] Oguzhan Tas and Farzad Kiyani. 2007. A survey automatic text summarization. *PressAcad. Proced.* 5, 1 (2007), 205–213.
[166] Maartje ter Hoeve, Julia Kiseleva, and Maarten de Rijke. 2020. What makes a good summary? Reconsidering the focus of automatic summarization. *arXiv preprint arXiv:2012.07619*.
[167] Akanksha Tiwari, Christian Von Der Weth, and Mohan S. Kankanhalli. 2018. Multimodal multiplatform social media event summarization. *ACM Trans. Multim. Comput., Commun. Applic.* 14, 2s (2018), 1–23.
[168] Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. Multi-modal summarization of key events and top players in sports tournament videos. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'11)*. IEEE, 471–478.
[169] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. 2021. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*.

[170] Chih-Fong Tsai, Kuanchin Chen, Ya-Han Hu, and Wei-Kai Chen. 2020. Improving text summarization of online hotel reviews with review helpfulness and sentiment. *Tour. Manag.* 80 (2020), 104122.

[171] Naushad UzZaman, Jeffrey P. Bigham, and James F. Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*. ACM, 43–52.

[172] Oleg Vasilyev and John Bohannon. 2021. Is human scoring the best criteria for summary evaluation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP'21*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2021.findings-acl.192

[173] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 5998–6008.

[174] Yash Verma, Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Dwaipayan Roy. 2022. MAKED: Multi-lingual automatic keyword extraction dataset. In *Proceedings of the 13th Language Resources and Evaluation Conference*. 6170–6179.

[175] Mila Vulchanova, Valentin Vulchanov, Isabella Fritz, and Evelyn A. Milburn. 2019. Language and perception: Introduction to the special issue "Speakers and listeners in the visual world". *Journal of Cultural Cognitive Science* 3 (2019), 103–112.
Language and perception: Introduction to the special issue speakers and listeners in the visual world. *J. Cultur. Cogn. Sci.* (2019), 1–10.

[176] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.

[177] Nancy X. R. Wang, Ali Farhadi, Rajesh P. N. Rao, and Bingni W. Brunton. 2018. AJILE movement prediction: Multimodal deep learning for natural human neural recordings and video. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[178] Shize Xu, Liang Kong, and Yan Zhang. 2013. A cross-media evolutionary timeline generation framework based on iterative recommendation. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*. 73–80.

[179] Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. A deep multi-level attentive network for multimodal sentiment analysis. *arXiv preprint arXiv:2012.08256*.

[180] Rui Yan, Xiaojun Wan, Mirella Lapata, Wayne Xin Zhao, Pu-Jen Cheng, and Xiaoming Li. 2012. Visualizing timelines: Evolutionary summarization via iterative reinforcement between text and image streams. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 275–284.

[181] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowl. Inf. Syst.* 53, 2 (2017), 297–336.

[182] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Computat. Ling.* 2 (2014), 67–78.

[183] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.306/.

[184] Naitong Yu, Minlie Huang, Yuanyuan Shi, and Xiaoyan Zhu. 2016. Product review summarization by exploiting phrase properties. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*. 1113–1124.

[185] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

[186] Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. 2016. Is a picture worth a thousand words? A Deep Multi-modal Fusion Architecture for Product Classification in e-commerce. *arXiv preprint arXiv:1611.09534*.

[187] Jiaming Zhan, Han Tong Loh, and Ying Liu. 2009. Gather customer concerns from online product reviews–A text summarization approach. *Expert Syst. Applic.* 36, 2 (2009), 2107–2115.

[188] Luming Zhang, Yue Gao, Chao Zhang, Hanwang Zhang, Qi Tian, and Roger Zimmermann. 2014. Perception-guided multimodal feature fusion for photo aesthetics assessment. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 237–246.

[189] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5674–5681.

[190] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=SkeHuCVFDr.

[191] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 563–578. DOI: https://doi.org/10.18653/v1/D19-1053

[192] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2017. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *arXiv preprint arXiv:1801.00054*.

[193] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13041–13049.

[194] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 4154–4164.

[195] Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9749–9756.

[196] Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9749–9756.

[197] Hai Zhuge. 2016. *Multi-dimensional Summarization in Cyber-physical Society*. Morgan Kaufmann.

[198] Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikolaos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos. 2017. COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP J. Image Vid. Process.* 2017, 1 (2017), 1–24.

Received 5 January 2021; revised 3 December 2022; accepted 12 December 2022