MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos

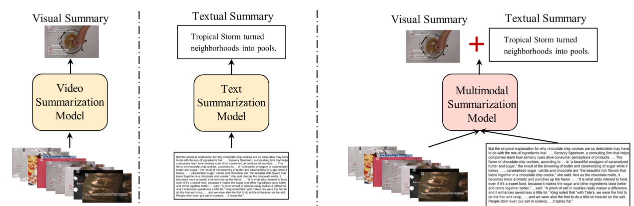
Jielin Qiu[1,2], Jiacheng Zhu[3], William Han[1], Aditesh Kumar[1], Karthik Mittal[1], Claire Jin[1], Zhengyuan Yang[2], Linjie Li[2], Jianfeng Wang[2], Ding Zhao[1], Bo Li[4], Lijuan Wang[2]
[1]Carnegie Mellon University, [2]Microsoft Azure AI, [3]MIT CSAIL, [4]University of Chicago
{jielinq,wjhan,dingzhao}@andrew.cmu.edu, zjc@mit.edu lbo@illinois.edu
{zhengyang,lindsey.li,jianfw,lijuanw}@microsoft.com

Abstract

Multimodal summarization with multimodal output (MSMO) has emerged as a promising research direction. Nonetheless, numerous limitations exist within existing public MSMO datasets, including insufficient maintenance, data inaccessibility, limited size, and the absence of proper categorization, which pose significant challenges. To address these challenges and provide a comprehensive dataset for this new direction, we have meticulously curated the ***MMSum*** *dataset. Our new dataset features (1) Human-validated summaries for both video and textual content, providing superior human instruction and labels for multimodal learning. (2) Comprehensively and meticulously arranged categorization, spanning 17 principal categories and 170 subcategories to encapsulate a diverse array of real-world scenarios. (3) Benchmark tests performed on the proposed dataset to assess various tasks and methods, including video summarization, text summarization, and multimodal summarization. To champion accessibility and collaboration, we released the* ***MMSum*** *dataset and the data collection tool as fully open-source resources, fostering transparency and accelerating future developments, at* https://mmsum-dataset.github.io/.

1. Introduction

Multimodal summarization with multimodal output (MSMO) is an emerging research topic spurred by advancements in multimodal learning [10, 30, 36, 62, 130] and the increasing demand for real-world applications such as medical reporting [49], educational materials [81], and social behavior analysis [51]. Most MSMO studies focus on video data and text data, aiming to select the most informative visual keyframes and condense the text content into key points. In this study, we focus on MSMO, which integrates both visual and textual information to provide users with comprehensive and representative summaries to enhance user experience [19, 43, 130].



(a) Video summarization. (b) Text summarization. (c) MSMO: Multimodal summarization with multimodal output.

Figure 1. Task comparison of traditional video summarization, text summarization, and MSMO tasks.

Despite the respective accomplishments of conventional unimodal summarization techniques on video data [35, 69, 83, 118, 123, 127, 131] and text data [14, 55, 60, 61, 125], multimodal summarization continues to pose challenges due to a number of complexities. (1) The intricate nature of multimodal learning necessitates an algorithm capable of exploiting correlated information across different modalities, (2) There is a scarcity of appropriate multimodal datasets that reliably exhibit cross-modal correlations across diverse categories, and (3) There exists a gap in comprehensive evaluation protocols that accurately reflect the efficacy of MSMO methods in terms of their performance on both intermediate interpretations and downstream tasks.

Merging existing video and text datasets appears to be a feasible approach. However, assuring the presence of cross-modal correlations proves challenging [62], not to mention the absence of necessary human verification [67], a vital element in machine learning research. Furthermore, the existing datasets pose several issues, such as inadequate maintenance leading to data unavailability, limited size, and lack of categorization. To address these concerns and offer a comprehensive dataset for this area of study, we have undertaken the task of collecting a new dataset, named **MMSum**. Our contributions are summarised as follows:

- **A new MSMO dataset** Introducing MMSum, our newly curated MSMO dataset, specifically designed to

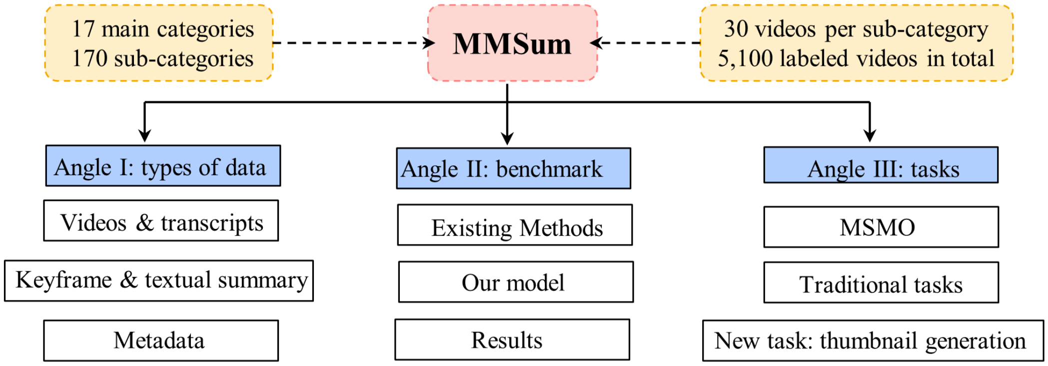


Figure 2. The design of the proposed MMSum dataset is driven by research and application needs.

cater to a wide range of tasks, with a particular emphasis on MSMO. This extensive dataset offers abundant information that serves as solid support for various research endeavors.

- **Diverse categorization** Within the MMSum dataset, we have meticulously gathered videos spanning 17 primary categories. Each of these main categories further comprises 10 distinct subcategories, culminating in a grand total of 170 subcategories. This comprehensive categorization ensures that the MMSum dataset is exceptionally representative and encompasses a wide range of content.
- **New benchmark** Across a diverse array of tasks, our results can be regarded as a benchmark on this novel real-world dataset.
- **Accessibility** We will open-source the MMSum dataset and the corresponding data collection tool with CC BY-NC-SA License.

2. Related Work

Unimodal Summarization typically comprises video summarization and text summarization. Video summarization involves extracting key moments that summarize the content of a video by selecting the most informative and essential parts. Traditional video summarization methods primarily rely on visual information. However, recent advancements have introduced category-driven or supervised approaches that generate video summaries by incorporating video-level labels, thereby enhancing the summarization process [25, 63, 94, 107, 127, 128]. Text Summarization involves processing textual metadata, such as documents, articles, tweets, and more, as input, and generating concise textual summaries. The quality of generated summaries has recently been significant improved through fine-tuning pre-trained language models [48, 121].

Multimodal Summarization explored multiple modalities for summary generation. [19, 66, 105, 112] learned the relevance or mapping in the latent space between different modalities. In addition to only generating visual summaries, [3, 42, 130] generated textual summaries by taking audio, transcripts, or documents as input along with videos or images, using seq2seq model [96] or attention mechanism [5]. The methods above explored using multiple modalities' information to generate single modality output, either textual or visual summary. Recent trends on the MSMO task have also drawn much attention [19, 20, 29, 57, 77, 78, 99, 119, 122, 130]. Specifically, [99] summarized a video and text document into a cover frame and a one-sentence summary. The most significant difference between multimodal summarization and MSMO lies in the inclusion of multiple modalities in the output. (More related work can be found in Appendix G.)

3. Angle I: Types of data

3.1. Data Collection

In light of the aforementioned challenges inherent in the existing MSMO datasets, we propose a novel dataset named MMSum to address these issues comprehensively and effectively. Our approach involved the collection of a multimodal dataset, primarily sourced from a diverse range of untrimmed videos from YouTube. The collected dataset comprises a rich set of information, including video files and transcripts, accompanied by corresponding video metadata. Additionally, temporal boundaries were meticulously recorded for each segment within the videos. Furthermore, for each segment, we obtained both video summaries and text summaries. It is worth noting that these summaries were directly provided by the authors of the respective videos, ensuring their authenticity and reliability. Moreover, the dataset incorporates comprehensive video metadata, such as titles, authors, URLs, categories, subcategories, and so on. By gathering this diverse range of multimodal data and leveraging the ground-truth video and text summaries provided by the original content creators, we aim to create a valuable and reliable resource.

Table 1. Comparison of the modality of different summarization tasks and datasets. Difference between traditional multimodal summarization and MSMO: traditional multimodal summarization still outputs a single-modality summary, while MSMO outputs both modalities' summaries. Public Availability means whether the data is still publicly available and valid. Structural Summaries means available summaries of each segment, not just for the whole video.

Tasks	Datasets	Input		Output		Public Availability	Categorization	Structural Summaries
		Visual	**Textual**	**Visual**	**Textual**			
Video	TVSum [95]	✓	✗	✓	✗	✓	✗	✓
	SumMe [23]	✓	✗	✓	✗	✓	✗	✓
	VSUMM [16]	✓	✗	✓	✗	✓	✗	✓
Textual	X-Sum [64]	✗	✓	✗	✓	✓	✗	✗
	Pubmed [90]	✗	✓	✗	✓	✓	✗	✗
Multimodal	How2 [86]	✓	✓	✓	✗	✓	✗	✗
	AVIATE [3]	✓	✓	✗	✓	✓	✗	✗
	Daily Mail [130]	✓	✓	✗	✗	✓	✗	✗
MSMO	VMSMO [57]	✓	✓	✓	✓	✗	✗	✗
	MM-AVS [19]	✓	✓	✓	✓	✓	✗	✗
	MMSum (Ours)	✓	✓	✓	✓	✓	✓	✓

Fidelity Given the limited availability of fully annotated videos with complete and non-missing video summaries and text summaries, we resorted to a manual collection of videos that satisfied all the specified criteria. The meticulous nature of this process ensured that only videos meeting the stringent requirements were included in the dataset. To illustrate the disparities between different tasks and datasets in terms of modalities, we provide a comprehensive comparison in Table 1. Traditional video or text summarization datasets typically encompass either visual or textual information exclusively. While there are datasets available for traditional multimodal summarization, where multiple modalities are used as input, they still produce single-modality summaries. In contrast, the MSMO dataset holds significant value in real-world applications, as it requires multimodal inputs and provides summaries containing both visual and textual elements. Consequently, the collection process for this dataset necessitates acquiring all the requisite information, resulting in a time-consuming endeavor.

Human Verification Notably, every video in the MMSum dataset undergoes manual verification to ensure high-quality data that fulfills all the specified requirements. For the fidelity verification process, five human experts (3 male and 2 female) each spent 30 days watching the collected videos, understanding the content, and verifying the annotations. The annotators were instructed to pay specific attention to the quality of segmentation boundaries, visual keyframes, and textual summaries. The pre-filtered size of the dataset is 6,800 (40 videos per subcategory). After manual verification and filtering, only 30 of 40 are preserved to ensure the quality, resulting in the current size of 5,100 (30 videos per subcategory).

Diversity During the dataset creation process, we extensively examined existing video datasets such as [53, 129] for reference. Subsequently, we carefully selected 17 main categories to ensure comprehensive coverage of diverse topics. These main categories encompass a wide range of subjects, including *animals, education, health, travel, movies, cooking, job, electronics, art, personal style, clothes, sports, house, food, holiday, transportation*, and *hobbies*. Each main category is further divided into 10 subcategories based on the popularity of Wikipedia, resulting in a total of 170 subcategories. To illustrate the subcategories associated with each main category, please refer to Figure 3 and Table 6 (in the Appendix). For a more detailed view, a high-resolution version of Figure 3 can be found in Appendix B. To ensure the dataset's representativeness and practicality, we imposed certain criteria for video inclusion. Specifically, we only collected videos that were longer than 1 minute in duration while also ensuring that the maximum video duration did not exceed 120 minutes. Adhering to these guidelines allows a balance between capturing sufficient content in each video and preventing excessively lengthy videos from dominating the dataset. In total, our dataset comprises 170 subcategories and a grand total of 5,100 videos, all carefully selected to encompass a wide range of topics and characteristics.

3.2. Statistics of the Dataset

Figure 4 presents a comprehensive analysis of the MMSum dataset's statistics. Figure 4(a) delves into the distribution of video durations, revealing the average duration spans approximately 15 minutes. In Figure 4(b), we show the distribution of the number of segments per video. The graph in Figure 4(c) captures the distribution of segment durations, showcasing an intriguing resemblance to the Gaussian distribution with an approximate mean of 80 seconds. Figure 4(d) shows the distribution of the number of words per sentence.

Figure 3. The 17 main categories of the MMSum dataset, where each main category contains 10 subcategories, resulting in 170 subcategories in total. More details are listed in Table 6.

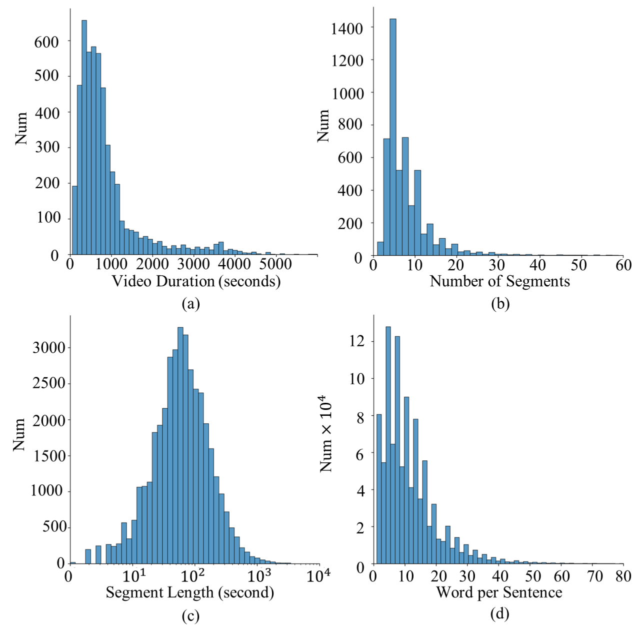


Figure 4. The statistics of the MMSum dataset, which show the distribution of (a) video duration; (b) number of segments per video; (c) segment duration; (d) number of words per sentence.

Table 2. Comparison with existing video summarization and multimodal summarization datasets.

	SumMe [23]	TVSum [95]	OVP [4]	CNN [20]	Daily Mail [20]	Ours
Source	YouTube	YouTube	YouTube	News	News	YouTube
Number of Data	25	50	50	203	1,970	**5,100**
Total Video Duration (Hours)	1.0	3.5	1.3	7.1	44.2	**1229.9**
Average Video Duration (mins)	2.4	3.9	1.6	2.1	1.4	**14.5**
Max Video Duration (mins)	5.4	10.8	3.5	6.9	4.8	**115.4**
Min Video Duration (mins)	0.5	1.4	0.8	0.3	0.4	1.0
Total Number of Text Tokens	–	–	–	0.2M	1.3M	**11.2M**
Avg. Keyframes per video	44	70	9.6	7.1	2.9	7.8
Avg. Text Summary Length	–	–	–	29.7	59.6	21.69
Number of Classes	25	10	7	–	–	**170**

3.3. Comparison with Existing Datasets

Table 2 presents a comparison between our MMSum dataset and existing video datasets. In contrast to standard video summarization datasets such as SumMe [23], TVSum [95], and OVP [4], our dataset, MMSum, stands out in several aspects. Firstly, the existing datasets lack textual data, whereas MMSum incorporates both video and textual information. Additionally, while the number of videos in SumMe, TVSum, and OVP is under 50, MMSum contains a substantial collection of 5,100 videos. Furthermore, the average duration of the videos in the aforementioned datasets is less than 4 minutes, whereas the videos in MMSum have an average duration of 14.5 minutes. Moreover, MMSum provides a significantly larger number of segments/keyframes per video compared to these standard datasets, making it more suitable for real-world applications. Comparing MMSum with other MSMO datasets like CNN and Daily Mail [20], we find that our dataset first surpasses them in terms of the number of videos. Furthermore, CNN and Daily Mail datasets were not curated based on specific classes; instead, the data was randomly downloaded, resulting in a lack of representativeness. In contrast, MMSum was carefully designed with 17 main categories and 170 subcategories, making it highly representative and practical. Although there are other MSMO datasets like VMSMO [57], we did not include them in the comparison table due to a large portion of the video links no longer be valid. Therefore, MMSum stands out as a comprehensive and reliable dataset for multimodal summarization tasks. The key distinguishing features of MMSum can be summarized as follows:

- MMSum offers an extensive and large-scale dataset, comprising an impressive collection of 5,100 human-annotated videos.
- The dataset showcases a remarkable range of untrimmed videos, varying in duration from concise 1-minute clips to extensive recordings spanning up to 115 minutes. This diversity allows for a comprehensive exploration of different video lengths and content complexities.
- MMSum's strength lies in its meticulously crafted main category and subcategory groups, which exhibit an exceptional level of richness and granularity. With a keen focus on real-world applicability, these categories are thoughtfully designed to encapsulate the diverse facets and contexts of video data, ensuring relevance across a wide array of domains.
- To guarantee the highest quality and integrity of the dataset, MMSum undergoes rigorous manual verification. This meticulous process ensures that all modalities and information within the dataset are accurately annotated and readily accessible.

4. Angle II: Benchmark

4.1. Problem Formulation

The formulation of the MSMO task can be expressed as follows. A video and its corresponding transcripts are denoted as a pair (V, X). The video input, represented by V, consists of a sequence of frames: $V = (v_1, v_2, \ldots, v_N)$. The corresponding transcripts, denoted as X, are a sequence of sentences: $X = (x_1, x_2, \ldots, x_M)$,. Note that M may not equal N due to one sentence per frame is not guaranteed in real-world videos. It is assumed that each video has a sequence of ground-truth textual summary, denoted as $Y = (y_1, y_2, \ldots, y_L)$, and a sequence of ground-truth keyframe represented by $P = (p_1, p_2, \ldots, p_L)$, where L is the number of segments. The objective of the MSMO task is to generate textual summaries $\widehat{Y}$ that capture the main points of the video, and select keyframes $\widehat{P}$ from V to be the visual summaries.

4.2. Existing Methods

In order to conduct a thorough performance evaluation, we selected a set of established methods as our baselines. These baselines are chosen based on the public availability of official implementations, ensuring reliable and reproducible results. The selected baseline methods encompass:

- For Video Summarization: Uniform Sampling [33], K-means Clustering [26], VSUMM [16], and Keyframe Extraction [33].
- For Text Summarization: BERT2BERT [100], BART [41] (BART-large-CNN and BART-large-XSUM), Distilbart [91], T5 [80], Pegasus [117], and LED [6].

More details of the baselines within the benchmark can be found in Appendix E. However, due to the absence of publicly available implementations for MSMO methods in the existing literature, there are no suitable methods that can be used as MSMO baselines.

4.3. Our Method

To solve the problem mentioned above and provide a MSMO baseline for the collected MMSum dataset, we propose a novel and practical approach to augment the MSMO baseline. Our method, which we have made accessible on our website, comprises two modules: segmentation and summarization. Our model is depicted in Figure 5.

Segmentation Module The primary objective of the segmentation module is to partition a given video into smaller segments based on the underlying content. This module operates by leveraging the entire transcript associated with the video, employing a contextual understanding of the text. For the segmentation module, we adopted a hierarchical BERT architecture, which has demonstrated state-of-the-art performance [50]. It comprises two transformer encoders. The first encoder focuses on sentence-level encoding, while the second encoder handles paragraph-level encoding. The first encoder encodes each sentence independently using $\text{BERT}_{\text{LARGE}}$ and then feeds the encoded embeddings into the second encoder. Notably, all sequences commence with a special token [CLS] to facilitate encoding at the sentence level. If a segmentation decision is made at the sentence level, the [CLS] token is utilized as input for the second encoder, which enables inter-sentence relationships to be captured through cross-attention mechanisms. This enables a cohesive representation of the entire transcript, taking into account the contextual dependencies between sentences.

Summarization Module Upon segmenting the video, each video segment becomes the input to the summarization module. In line with the model architecture proposed in [37], we construct our summarization module. The summarization module incorporates three main encoders: a frame encoder, a video encoder, and a text encoder. These encoders are responsible for processing the video frames, video content, and corresponding text, respectively, to extract relevant feature representations. Once the features have been extracted, multi-head attention is employed to fuse the learned features from the different encoders, which allows for the integration of information across the modalities, enabling a holistic understanding of the video and its textual content. Following the fusion of features, a score calculation step is performed to select the keyframe,

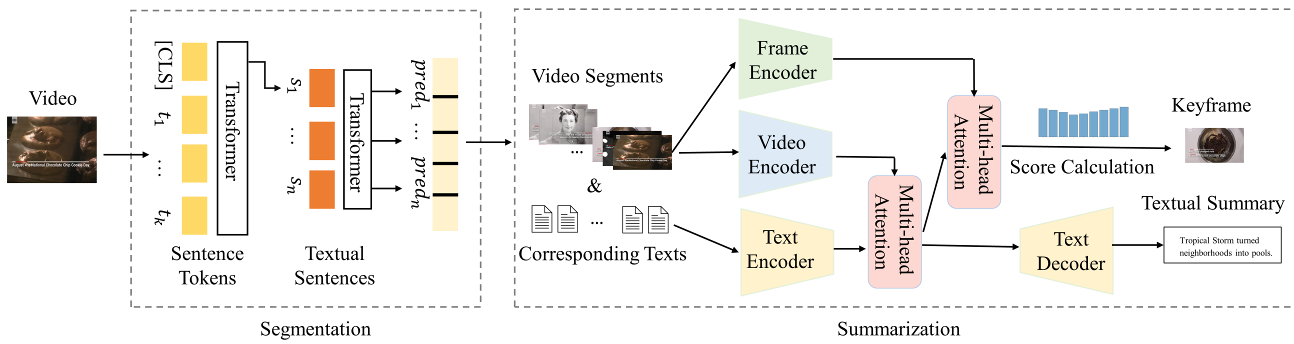


Figure 5. Our model comprises two modules: the segmentation module and the summarization module.

identifying the most salient frame within each video segment. Additionally, a text decoder is utilized to generate the textual summary, leveraging the extracted features and the fused representations. Considering our primary focus on providing a benchmark in this work, we have included model details in Appendix D due to page limit.

5. Angle III: Tasks and Results

5.1. Types of tasks

Within our dataset, a wealth of information is available, enabling the exploration of various downstream tasks. These tasks encompass video summarization (VS), text summarization (TS), and multimodal video summarization with multimodal output (MSMO). To provide a comprehensive understanding of each task and highlight their distinctions, we have compiled detailed descriptions and comparisons in Appendix C. For the train/val/test split, since our dataset is already randomly collected from YouTube, we designate the last 30% of videos within each subcategory (indexed 21-29) as the testing set. The remaining videos are then assigned to the training set (indexed 00-20) in each subcategory. More results are shown in Appendix F.

5.2. Evaluation of Traditional Tasks

Video Summarization Evaluation The quality of the chosen keyframe is evaluated by Root Mean Squared Error (RMSE), Structural Similarity Index (SSIM), Signal reconstruction error ratio (SRE), and Spectral angle mapper (SAM), between image references and the extracted video frames [59]. In addition, we also adopted precision, recall, and F1 score based on SSIM for evaluation.

Text Summarization Evaluation The quality of generated textual summary is evaluated by standard evaluation metrics, including BLEU [68], METEOR [17], ROUGE-L [45], CIDEr [102], and BertScore [120], following previous works [13, 57, 89]. ROUGE-1, ROUGE-2, and ROUGE-L refer to the overlap of unigram, bigrams, and the longest common subsequence between the decoded summary and the reference, respectively [45].

5.3. Results and Discussion

Supervised methods outperform unsupervised methods on video summarization In our video summarization study, we have chosen the following methods as our baseline comparisons: Uniform Sampling [33], K-means Clustering [26], and VSUMM [16]. The results, presented in Table 3, are under various evaluation metrics. For RMSE and SRE, lower values indicate better performance, whereas, for the remaining metrics, higher values are desirable. From Table 3, we can observe that VSUMM showcases the strongest performance among the baseline methods, yet it still falls short compared to our proposed method. But we can conclude that supervised methods outperform unsupervised methods.

Pretrained large language models can still do well in text summarization In the context of textual summarization, we have considered a set of representative models as our baseline comparisons: BERT2BERT [100], BART [41] (including BART-large-CNN and BART-large-XSUM), Distilbart [91], T5 [80], Pegasus [117], and Longformer Encoder-Decoder (LED) [6]. The performance of these models is summarized in Table 4. Among the baselines, T5, BART-large-XSUM, BART-large-CNN, and BERT2BERT exhibit superior performance, with T5 demonstrating relatively better results across various text evaluation metrics. In addition, the ROUGE score may not effectively capture performance differences compared to other evaluation metrics, because ROUGE does not take into account the semantic meaning and the factual accuracy of the summaries.

MSMO results may depend on segmentation results and summarization methods In the field of MSMO, we encountered limitations in accessing the codebases of existing works such as [10, 19, 20, 30, 113, 130]. Therefore, we independently implemented several baselines to evaluate their performance on the MMSum dataset. For this purpose, we utilized LGSS as the segmentation backbone, VSUMM as the video summarizer, and selected text summarizers that

Table 3. Comparison of video summarization results (whole-video setting and segment-level setting).

Setting	Model	RMSE ↓	PSNR ↑	SSIM ↑	SRE ↓	Precision ↑	Recall ↑	F1 Score ↑
Whole-video	Uniform [33]	0.479	4.044	0.076	49.808	0.077	0.100	0.049
	K-means [26]	0.348	8.234	0.055	46.438	0.072	0.182	0.103
	VSUMM [16]	0.279	9.226	0.053	44.862	0.054	0.259	0.088
	Ours	**0.112**	**25.280**	**0.697**	**23.550**	**0.320**	**0.290**	**0.321**
Segment-level	Uniform [33]	0.237	6.307	0.085	42.495	0.186	0.179	0.105
	K-means [26]	0.167	10.123	0.144	46.533	0.123	0.172	0.143
	VSUMM [16]	0.122	18.818	0.258	41.601	0.160	0.207	0.171
	Ours	**0.091**	**36.370**	**0.698**	**23.430**	**0.333**	**0.275**	**0.255**

Table 4. Comparison of textual summarization results (whole-video setting and segment-level setting).

Setting	Model	BLEU-1 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	METEOR ↑	CIDEr ↑	SPICE ↑	BertScore ↑
Whole-video	BERT2BERT [100]	22.59	3.75	0.45	3.41	5.65	1.76	2.91	71.12
	BART-large-CNN [41]	29.17	3.19	0.51	3.04	2.99	2.28	11.27	68.84
	BART-large-XSUM [41]	30.91	3.83	0.57	3.59	3.99	2.56	3.71	69.56
	Distilbart [91]	26.46	3.87	3.87	0.47	3.59	2.25	4.16	69.37
	T5 [80]	25.39	3.51	0.43	3.21	4.51	1.97	5.66	70.38
	Pegasus [117]	26.73	3.75	0.52	3.40	4.52	2.38	7.82	68.92
	LED [6]	26.47	3.81	0.25	3.51	3.45	1.78	6.72	68.45
	Ours	**32.61**	**9.41**	**2.86**	**9.15**	4.01	4.01	10.11	**74.46**
Segment-level	BERT2BERT [100]	13.58	4.70	1.95	4.53	28.59	11.73	10.13	71.76
	BART-large-CNN [41]	22.79	6.45	2.46	6.32	26.21	20.64	10.13	71.44
	BART-large-XSUM [41]	20.89	7.31	2.77	7.13	29.36	20.90	10.20	71.42
	Distilbart [91]	14.77	1.95	0.15	1.87	23.52	11.83	10.53	66.46
	T5 [80]	16.48	6.17	3.03	5.99	28.22	20.96	10.35	71.95
	Pegasus [117]	16.17	3.41	0.96	3.29	29.82	17.26	10.39	67.81
	LED [6]	16.03	3.80	0.60	3.64	29.81	15.85	10.99	68.46
	Ours	**23.36**	**13.61**	**4.58**	**13.24**	**30.01**	**21.06**	10.28	**85.19**

exhibited the best performance in text summarization. The results are presented in Table 5. Based on the findings, it is evident that the aforementioned combination approaches still fall short in comparison to our proposed method. This also indicates that the accuracy of temporal segmentation is crucial prior to generating summaries, highlighting it as a critical step and task preceding MSMO.

5.4. Thumbnail Generation

One direct and practical application of the MSMO task is to automatically generate thumbnails for a given video, which has become increasingly valuable in various real-world applications. With the exponential growth of online videos, effective and efficient methods are required to extract visually appealing and informative thumbnail representations. In addition, many author-generated thumbnails involve words or titles that describe the whole video to attract more users. In the context of online platforms, such as video-sharing websites or social media platforms, compelling thumbnails can significantly impact user engagement, content discoverability, and overall user experience. The benefits of automated thumbnail generation extend beyond user engagement and content discoverability. In e-commerce, for instance, thumbnails can play a vital role in attracting potential buyers by effectively showcasing products or services. Similarly, in video editing workflows, quick and accurate thumbnail generation can aid content creators in managing and organizing large video libraries efficiently.

In our setting, we take advantage of the results by MSMO, which contains both visual summary and text summary, and combine them to generate thumbnails for a given video. In summary, the selected keyframes and generated textual summaries from the MSMO task are subsequently utilized to create the thumbnail. To ensure an aesthetically pleasing appearance, we randomly sample from a corpus of fonts from Google Fonts and font sizes to utilize in the generated thumbnails. Moreover, a random set of coordinates on the selected keyframe is sampled for the placement of the text. Finally, the text is pasted onto the keyframe from the outputted set of coordinates to complete thumbnail generation.

More specifically, the font is randomly selected from 100 fonts, and the size of the font varies by 175 font sizes. Here we list 20 examples of fonts we used in our experiments: [Roboto, Open Sans, Lato, Montserrat, Raleway, Oswald, Source Sans Pro, Poppins, Noto Sans, Roboto Slab, Merriweather, Ubuntu, PT Sans, Playfair Display, Fira Sans, Nunito, Roboto Condensed, Zilla Slab, Arvo, Muli]. We randomly select one font and a random font size. Given

Table 5. Comparison of MSMO results.

Methods	Text					Video				
	BLEU ↑	METEOR ↑	CIDEr ↑	SPICE ↑	BertScore ↑	PSNR ↑	SSIM ↑	Precision ↑	Recall ↑	F1 Score ↑
LGSS + VSUMM + T5	27.35	24.32	3.94	5.57	62.77	16.234	0.198	0.143	0.152	0.147
LGSS + VSUMM + BART-large-XSUM	24.83	24.12	3.97	8.86	39.20	16.234	0.198	0.143	0.152	0.147
LGSS + VSUMM + BERT2BERT	13.26	24.83	3.68	9.23	64.34	16.234	0.198	0.143	0.152	0.147
LGSS + VSUMM + BART-large-CNN	24.93	28.61	3.78	9.84	64.44	16.234	0.198	0.143	0.152	0.147
Ours	**33.36**	**30.31**	**4.06**	**10.28**	**85.19**	**36.370**	**0.298**	**0.233**	**0.275**	**0.155**

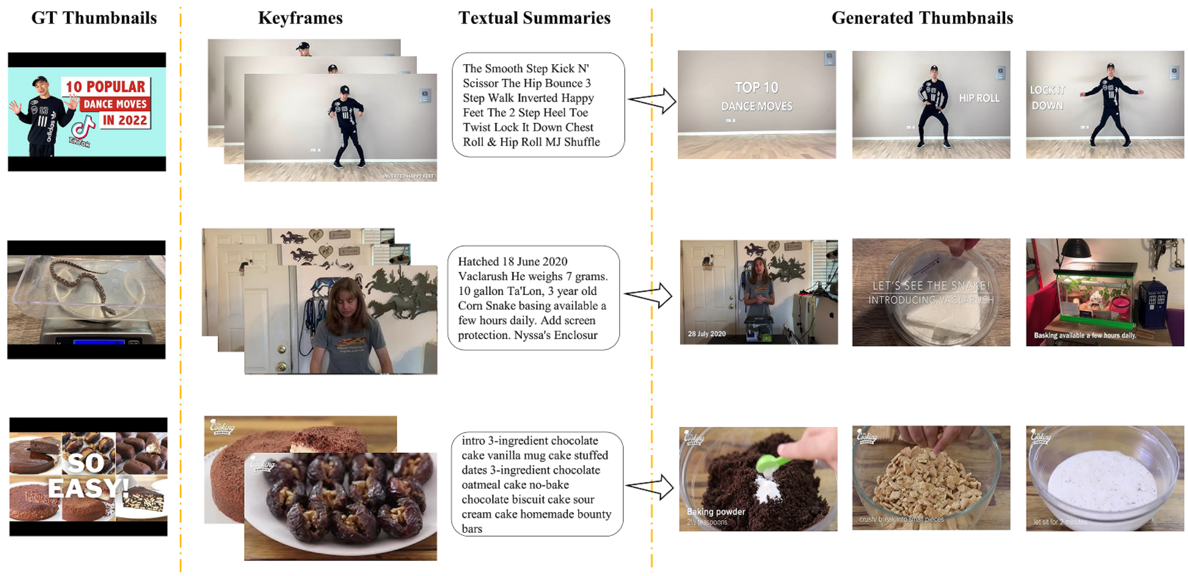


Figure 6. Comparison of GT thumbnails and our generated ones.

the image size of the selected keyframes, we also randomly select coordinates for where the text should be pasted onto the selected keyframes. We then paste the generated textual summary, which is modified by the randomly selected font and font size, onto the selected keyframes. Some examples are shown in Figure 6. More results can be found in Appendix H.

Limitations and Future Work Directions The lack of publicly available MSMO baselines in existing literature underscores a significant gap, emphasizing the need for future efforts in this area. Advancing the field requires tackling the complex task of creating a diverse and extensive collection of baselines.

Despite the progress made in automated thumbnail generation, challenges remain. These include enhancing the accuracy of thumbnail selection, accommodating various video genres and content types, and taking into account user preferences and context-specific requirements.

Moreover, addressing ethical concerns related to potential biases, representation, and content moderation is crucial to ensuring fair and inclusive thumbnail generation. Exploring new quantitative evaluation metrics for the thumbnail generation task could also pave the way for valuable advancements in this domain.

6. Conclusion

In this research, our main goal was to overcome the limitations of existing MSMO datasets by creating a comprehensive dataset called MMSum. MMSum was meticulously curated to ensure top-notch quality of MSMO data, making it a valuable resource for tasks like video summarization, text summarization, and multimodal summarization. Additionally, we introduced a novel benchmark based on the MMSum dataset. This benchmark enables researchers and practitioners to assess their algorithms and models across a range of tasks. Moreover, leveraging the results from MSMO, we introduced a new task: automatically generating thumbnails for videos. This innovation has the potential to significantly enhance user engagement, content discoverability, and overall user experience. We hope that our MMSum dataset can contribute to the advancement of research in the MSMO field.

Acknowledgement

We greatly appreciate the valuable feedback from Lei Li and Christos Faloutsos. This research is partially supported by Microsoft Azure AI, CMU Computer Science Department, and CMU GSA/Provost Conference Funding.

References

[1] Sathyanarayanan N. Aakur and Sudeep Sarkar. A perceptual prediction framework for self supervised event segmentation. *CVPR*, pages 1197–1206, 2019. 27

[2] Evlampios E. Apostolidis, E. Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and I. Patras. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109:1838–1863, 2021. 27

[3] Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *ArXiv*, abs/2105.09601, 2021. 2, 3

[4] Sandra Avila, Ana Paula Brandão Lopes, Antonio da Luz, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.*, 32:56–68, 2011. 4

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015. 2

[6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150, 2020. 5, 6, 7, 25, 26

[7] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Yue Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. Experience grounds language. In *EMNLP*, 2020. 27

[8] Tom B. Brown et al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 27

[9] Brandon Castellano. Intelligent scene cut detection and video splitting tool. `https://bcastell.com/projects/PySceneDetect/`, 2021. 24, 25

[10] Jingqiang Chen and Hai Zhuge. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *EMNLP*, pages 4046–4056, 2018. 1, 6, 27

[11] Shixing Chen, Xiaohan Nie, David D. Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. *CVPR*, pages 9791–9800, 2021. 27

[12] Xi Chen et al. Pali: A jointly-scaled multilingual language-image model. *ArXiv*, abs/2209.06794, 2022. 27

[13] Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. Iterative document representation learning towards summarization with polishing. In *EMNLP*, 2018. 6

[14] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. In *ACL*, pages 484–494, 2016. 1

[15] Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022. 27

[16] Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011. 3, 5, 6, 7, 24, 26

[17] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014. 6

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 22

[19] Xiyan Fu, Jun Wang, and Zhenglu Yang. Multi-modal summarization for video-containing documents. *ArXiv*, abs/2009.08018, 2020. 1, 2, 3, 6, 23, 27

[20] Xiyan Fu, Jun Wang, and Zhenglu Yang. Mm-avs: A full-scale dataset for multi-modal summarization. In *NAACL*, 2021. 2, 4, 6, 27

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64:86 – 92, 2018. 14

[22] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and Dhruv Kumar Mahajan. Large-scale weakly-supervised pre-training for video action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12038–12047, 2019. 22

[23] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014. 3, 4, 27

[24] William Han, Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Douglas Weber, Bo Li, and Ding Zhao. An empirical exploration of cross-domain alignment between language and electroencephalogram. *ArXiv*, abs/2208.06348, 2022. 27

[25] Li Haopeng, Ke Qiuhong, Gong Mingming, and Tom Drummond. Progressive video summarization via multimodal self-supervised learning. 2022. 2, 27

[26] J. A. Hartigan and M. Anthony. Wong. A k-means clustering algorithm. 1979. 5, 6, 7, 24, 26

[27] Ahmed Hassanien, Mohamed A. Elgharib, Ahmed A. S. Seleim, Mohamed Hefeeda, and Wojciech Matusik. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *ArXiv*, abs/1705.03281, 2017. 27

[28] Eman Hato and Matheel Emaduldeen Abdulmunem. Fast algorithm for video shot boundary detection using surf features. *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, pages 81–86, 2019. 27

[29] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14867–14878, 2023. 2

[30] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP*, pages 2352–2356, 2019. 1, 6, 27

[31] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *ACL*, 2019. 27

[32] Zhihu Huang and Jinsong Leng. Analysis of hu's moment invariants on image scaling and rotation. *2010 2nd International Conference on Computer Engineering and Technology*, 7:V7–476–V7–480, 2010. 24, 25

[33] Shruti Jadon and Mahmood Jasim. Video summarization using keyframe extraction and video skimming. *arXiv preprint arXiv:1910.04792*, 2019. 5, 6, 7, 24, 25, 26

[34] Vidhi Jain, Yixin Lin, Eric Undersander, Yonatan Bisk, and Akshara Rai. Transformers are adaptable task planners. *ArXiv*, abs/2207.02442, 2022. 27

[35] Hao Jiang and Yadong Mu. Joint video summarization and moment localization by cross-task sample transfer. In *CVPR*, pages 16388–16398, 2022. 1

[36] Aman Khullar and Udit Arora. Mast: Multimodal abstractive summarization with trimodal hierarchical attention. *arXiv preprint arXiv:2010.08021*, 2020. 1

[37] Mateusz Krubiński and Pavel Pecina. Mlask: Multimodal summarization of video-based news articles. In *Findings*, 2023. 5, 22, 23

[38] Hilde Kuehne, Alexander Richard, and Juergen Gall. A hybrid rnn-hmm approach for weakly supervised temporal action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:765–779, 2020. 27

[39] Colin S. Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017. 27

[40] Szer Ming Lee, John H. Xin, and Stephen Westland. Evaluation of image similarity by histogram intersection. *Color Research and Application*, 30:265–274, 2005. 24, 25

[41] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. 5, 6, 7, 25, 26

[42] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *EMNLP*, 2017. 2

[43] Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. Vmsmo: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406*, 2020. 1, 23

[44] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. 2022. 27

[45] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004. 6

[46] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Conference on Empirical Methods in Natural Language Processing*, 2020. 23

[47] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14:715–729, 2022. 27

[48] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *EMNLP*, pages 3730–3740, 2019. 2, 27

[49] Ziyi Liu, Jiaqi Zhang, Yongshuai Hou, Xinran Zhang, Ge Li, and Yang Xiang. Machine learning for multimodal electronic health records-based research: Challenges and perspectives, 2021. 1

[50] Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. Text segmentation by cross segment attention. In *EMNLP*, 2020. 5

[51] Anshu Malhotra and Rajni Jindal. Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. *EAI Endorsed Transactions on Pervasive Health and Technology*, 6(21), 2020. 1

[52] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2019. 22

[53] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 3

[54] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 22

[55] Derek Miller. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*, 2019. 1

[56] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. *ArXiv*, abs/2110.07342, 2022. 27

[57] Li Mingzhe, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. Vmsmo: Learning to generate multimodal summary for video-based news articles. *ArXiv*, abs/2010.05406, 2020. 2, 3, 4, 6

[58] Jean-Michel Morel and Guoshen Yu. Is the " scale invariant feature transform " (sift) really scale invariant ? 2010. 24, 26

[59] Markus U. Müller, Nikoo Ekhtiari, Rodrigo M. Almeida, and Christoph Rieke. Super-resolution of multispectral satellite images using convolutional neural networks. *ArXiv*, abs/2002.00580, 2020. 6

[60] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence

model for extractive summarization of documents. In *Thirty-first AAAI*, 2017. 1

[61] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *SIGNLL*, pages 280–290, 2016. 1

[62] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000, 2021. 1

[63] Medhini G. Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl;dw? summarizing instructional videos with task relevance & cross-modal saliency. *ArXiv*, abs/2208.06773, 2022. 2, 27

[64] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018. 3

[65] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL*, pages 1747–1759, 2018. 21, 27

[66] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Video summarization using deep semantic features. *ArXiv*, abs/1609.08758, 2016. 2

[67] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 1

[68] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[69] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Sumgraph: Video summarization via recursive graph modeling. In *ECCV*, pages 647–663. Springer, 2020. 1

[70] Viorica Patraucean et al. Perception test : A diagnostic benchmark for multimodal models. 2022. 27

[71] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2544, 2014. 27

[72] Jielin Qiu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, D. Zhao, and Hailin Jin. Liveseg: Unsupervised multimodal temporal segmentation of long livestream videos. *ArXiv*, abs/2210.05840, 2022. 27

[73] Jielin Qiu, William Jongwon Han, Jiacheng Zhu, Mengdi Xu, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Transfer knowledge from natural language to electrocardiography: Can we detect cardiovascular disease through language models? *ArXiv*, abs/2301.09017, 2023. 27

[74] Jielin Qiu, Peide Huang, Makiya Nakashima, Jae-Hyeok Lee, Jiacheng Zhu, W. H. Wilson Tang, Po-Heng Chen, Christopher Nguyen, Byung-Hak Kim, Debbie Kwon, Douglas Weber, Ding Zhao, and David Chen. Multimodal representation learning of cardiovascular magnetic resonance imaging. *ArXiv*, abs/2304.07675, 2023. 27

[75] Jielin Qiu, Mengdi Xu, William Jongwon Han, Seungwhan Moon, and Ding Zhao. Embodied executable policy learning with language-based scene summarization. *ArXiv*, abs/2306.05696, 2023. 27

[76] Jielin Qiu, Jiacheng Zhu, Shiqi Liu, William Jongwon Han, Jingqi Zhang, Chaojing Duan, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Converting ecg signals to images for efficient image-text retrieval via encoding. *ArXiv*, abs/2304.06286, 2023. 27

[77] Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. Mhms: Multimodal hierarchical multimedia summarization. *ArXiv*, abs/2204.03734, 2022. 2

[78] Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. Semantics-consistent cross-domain summarization via optimal transport alignment. *ArXiv*, abs/2210.04722, 2022. 2, 27

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 27

[80] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2019. 5, 6, 7, 25, 26

[81] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022. 1

[82] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. *CVPR*, pages 10143–10152, 2020. 24, 25, 27

[83] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the ECCV (ECCV)*, pages 347–363, 2018. 1

[84] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. `https://github.com/CompVis/stable-diffusion`. 27

[85] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2014. 22

[86] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. *ArXiv*, abs/1811.00347, 2018. 3

[87] M. Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. *CVPR*, pages 11220–11229, 2021. 27

[88] Christoph Schuhmann et al. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, abs/2111.02114, 2021. 27

[89] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointergenerator networks. In *ACL*, 2017. 6, 21, 27

[90] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Mag.*, 29:93–106, 2008. 3

[91] Sam Shleifer and Alexander M. Rush. Pre-trained summarization distillation. *ArXiv*, abs/2010.13002, 2020. 5, 6, 7, 25, 26

[92] Panagiotis Sidiropoulos, Vasileios Mezaris, Yiannis Kompatsiaris, Hugo Meinedo, Miguel M. F. Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21:1163–1177, 2011. 27

[93] Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Ndedi Monekosso, and Paolo Remagnino. Superframes, a temporal video segmentation. *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 566–571, 2018. 27

[94] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, pages 5179–5187, 2015. 2, 27

[95] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. *CVPR*, pages 5179–5187, 2015. 3, 4, 27

[96] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. 2

[97] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. Abstractive document summarization with a graph-based attentional neural model. In *ACL*, pages 1171–1181, 2017. 21, 27

[98] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019. 22

[99] Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. Tldw: Extreme multimodal summarisation of news videos. *ArXiv*, abs/2210.08481, 2022. 2

[100] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019. 5, 6, 7, 25, 26

[101] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 22

[102] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CVPR*, pages 4566–4575, 2015. 6

[103] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2740–2755, 2019. 27

[104] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *ECCV*, 2020. 27

[105] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. Video summarization via semantic attended networks. In *AAAI*, 2018. 2

[106] Yuxiang Wu and Baotian Hu. Learning to extract coherent summary via deep reinforcement learning. In *AAAI*, pages 5602–5609, 2018. 21, 27

[107] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. Convolutional hierarchical attention network for query-focused video summarization. *arXiv preprint arXiv:2002.03740*, 2020. 2, 27

[108] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*, 2020. 22

[109] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv*, abs/2303.11381, 2023. 27

[110] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. *ArXiv*, abs/2109.02401, 2021. 22, 23

[111] Lu Yuan et al. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432, 2021. 27

[112] Yitian Yuan, Tao Mei, Peng Cui, and Wenwu Zhu. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 29:226–237, 2019. 2

[113] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114, 2017. 6, 27

[114] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *CVPR*, 2022. 27

[115] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1:10–28, 1993. 24, 25

[116] Haoxin Zhang, Zhimin Li, and Qinglin Lu. Better learning shot boundary detection via multi-task. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 27

[117] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *ArXiv*, abs/1912.08777, 2019. 5, 6, 7, 25, 26

[118] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, pages 766–782. Springer, 2016. 1

[119] Litian Zhang, Xiaoming Zhang, Junshu Pan, and Feiran Huang. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *AAAI*, 2022. 2

[120] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2019. 6

[121] Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *ACL*, pages 5059–5069, 2019. 2, 27

[122] Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. Unims: A unified framework for multimodal summarization with knowledge distillation. In *AAAI*, 2021. 2

[123] Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2793–2801, 2021. 1

[124] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2933–2942, 2017. 27

[125] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*, 2020. 1

[126] Feng Zhou, Fernando De la Torre, and Jessica K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:582–596, 2013. 27

[127] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, pages 7582–7589, 2018. 1, 2, 27

[128] Kaiyang Zhou, T. Xiang, and A. Cavallaro. Video summarisation by classification with deep reinforcement learning. In *BMVC*, 2018. 2, 27

[129] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, pages 7590–7598, 2018. 3

[130] Junnan Zhu, Haoran Li, Tianshan Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. Msmo: Multimodal summarization with multimodal output. In *EMNLP*, 2018. 1, 2, 3, 6, 27

[131] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020. 1

[132] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *ICCV*, 2021. 27