
Multi-modal News Understanding with Professionally Labelled Videos (ReutersViLNews)

Shih-Han Chou^{*1,2}Matthew Kowal^{*1,4,7}Yasmin Niknam^{*1,3}Diana Moyano¹ Shayaan Mehdi¹ Richard Pito⁵ Cheng Zhang⁵ Ian Knopke⁵Sedef Akinli Kocak¹Leonid Sigal^{1,2,6}Yalda Mohsenzadeh^{1,3}Vector Institute¹, University of British Columbia², Western University³York University⁴, Reuters News Agency⁵, CIFAR⁶, Toyota Research Institute⁷

Abstract

While progress has been made in the domain of video-language understanding, current state-of-the-art algorithms are still limited in their ability to understand videos at high levels of abstraction, such as news-oriented videos. On the other hand, humans easily amalgamate information from video and language to infer information beyond what is visually observable in the pixels. An example of this is watching a news story, where the context of the event can play as big of a role in understanding the story as the event itself. Towards a solution for designing this ability in algorithms, here we present a large-scale analysis on an in-house dataset collected by the Reuters News Agency, called Reuters Video-Language News shortened to "ReutersViLNews" dataset which focuses on high-level video-language understanding with an emphasis on long-form news. The ReutersViLNews Dataset consists of long-form news videos collected and labeled by professionals in the news industry over several years and contains prominent news reporting from around the world. Each video involves a single story and contains action shots of the actual event, interviews with people associated with the event, footage from nearby areas, and more. ReutersViLNews dataset contains videos from seven subject categories: disaster, finance, entertainment, health, politics, sports, and miscellaneous with annotations from high-level to low-level, title caption, visual video description, high-level story description, keywords, and location. We first present a detailed analysis of the dataset statistics of ReutersViLNews compared to previous datasets. Then we benchmark state-of-the-art approaches for four different video-language tasks. The results suggest that news-oriented videos are a substantial challenge for current video-language understanding algorithms and we conclude by providing future directions in designing approaches to solve the ReutersViLNews dataset.

1 Introduction

The challenge of understanding video and language simultaneously remains an active area of research in machine learning [19, 20, 41, 9, 10, 27]. This task requires the ability to analyze and interpret both visual and linguistic information; several video understanding sub-tasks have been proposed, such as video captioning [3, 38, 37, 20, 19] (i.e., given a video, provide a caption describing the video), video question answering [50] (given a video and a question, provide the correct answer to the question),

^{*}Equal Contribution



Figure 1: **Dataset Examples.** Four example videos from the ReutersViLNews dataset. Frames are shown from four news categories with time increasing from left to right. Note the diversity of the events and visual information in each video clip. The stories covered are (i) Pogba an injury doubt for United ahead of Premier League clash with Arsenal; (ii) New Zealand’s COVID-19 cases hit record for second time this week; (iii) Britain will have “good set” of Brexit policies ready for June EU summit - minister; (iv) ‘Couldn’t think of a better place’ to announce baby: Prince Harry in Australia.

and text-video retrieval [28] (i.e., given a dataset of videos and sentences describing the videos, match the corresponding text-video pairs). While recent years have witnessed significant progress in the development of algorithms that can jointly learn from video and language data, many open challenges still exist, including the lack of high-quality benchmark datasets, the difficulty of modeling complex temporal dependencies [17], and the need for more effective and interpretable models. Therefore, the development of diverse and high-quality datasets must be prioritized to advance the state-of-the-art in video and language understanding. Several datasets [54, 36, 40, 6, 48, 51, 24, 16, 30] have been released to benchmark models on multiple video understanding tasks such as video classification [1, 31, 21, 43], video captioning [24, 38, 37, 20, 19], video paragraph generation [41, 27], and video-text retrieval [28, 45, 29]. The majority of these existing video benchmarks are either web-scraped from the internet [1, 24, 51], manually recorded and labeled by paid laypersons [24, 48] (e.g., Amazon Turk-like services), or relatively small in size [36, 6]. Moreover, the level of description of existing video-language datasets remains relatively low and focuses on describing events that visibly occur in the video.

Most of the approaches [21, 43, 42, 52] focus on the visual part to understand the content of the video. However, humans leverage visual, audio, linguistic, and even contextual information to understand videos at a higher level. For example, when watching a video about a natural disaster, a human understands the location of the event, the impact on surrounding areas, a sense of the level of damage, and other phenomena which are not visible by observing the pixels of the video. Despite this, there is a lack of video datasets that contain such abstract levels of textual annotations for video. Such a dataset would be beneficial as it would provide new technical challenges at higher levels of abstraction which would require the models to employ richer contextual information to solve the task at hand.

To spur such research, we propose a new video-language multimodal dataset: *the Reuters News agency Video-Language dataset shortened to “ReutersViLNews” dataset*, for a variety of video and video-language understanding tasks. ReutersViLNews contains news events videos, audio, and texts with professionally curated and labeled from around the world. It is collected by journalists from dozens of countries and contains rich and consistent labels generated by professionals in the news industry. To demonstrate the utility of the dataset, we benchmark various state-of-the-art algorithms for four different video-language understanding tasks and uncover interesting findings with respect to the type of challenges deep networks have with the dataset. Finally, we suggest future directions for addressing these challenges as well as open problems in the domain of video-language understanding.

Table 1: **Comparison of ReutersViLNews with other video language datasets.** #videos: number of videos in the corresponding dataset. Avg. len: average length of videos in seconds. Total len: Total length of videos in hours. Cap: captions. KeyW: keywords. C. Cap: closed caption.

Dataset	Source	Content	Video Statistics			Annotations			
			#videos	Avg. len	Total len	Cap.	Story	KeyW	C. Cap.
YouCook II [54]	YouTube	Cooking	2,000	316 sec	176 hrs	✓	✗	✗	✗
MPII-MD [36]	Films	Movie Scenes	68,337	3.9 sec	73.6 hrs	✓	✗	✗	✗
Charades [40]	Mech. Turk	Home Videos	9,848	30 sec	82 hrs	✓	✗	✗	✗
MSVD [6]	YouTube	Open	1,970	10 sec	5.3 hrs	✓	✗	✗	✗
MSR-VTT [48]	YouTube	Open	7,180	20 sec	41.2 hrs	✓	✗	✗	✗
VTW [51]	YouTube	Open	18,100	90 sec	213.2 hrs	✓	✗	✗	✗
ANet Caption [24]	YouTube	Open	20,000	180 sec	849 hrs	✓	✗	✗	✗
Howto100M [30]	YouTube	Instructional	1.2M	396 sec	134,472 hrs	✓	✗	✗	✗
VideoStory [16]	Social Media	Story Telling	20,147	70 sec	396 hrs	✓	✓	✗	✗
ReutersViLNews	Journalists	News	1,974	91.2 sec	50 hrs	✓	✓	✓	✓

This technical report can be summarized as follows: (i) We present an analysis on a new in-house dataset, ReutersViLNews: 1, 974 news story videos from 2018/04 to 2021/12 filmed and annotated by professional journalists and editors from the Reuters News agency, with the purpose of understanding current challenges in long-form news video understanding. The dataset contains videos from 200+ different global locations and spans seven subject categories: *disaster, finance, entertainment, health, politics, sports, and miscellaneous*. The language annotations of the dataset span various levels of abstraction (e.g., both visually observable actions and contextual background information) which include *title caption, visual video description, high-level story description, keywords, and location*. (ii) We then perform a comprehensive analysis of ReutersViLNews by comparing the statistics with other datasets and by evaluating several state-of-the-art deep models for four open problems in video-language understanding using the dataset: (1) video captioning, (2) video paragraph generation, (3) open-world video keyword generation, and (4) video-text retrieval. Our comprehensive experimental results suggest that video understanding of world events and news stories is a challenging domain.

2 Related Work

Video Description. The video description aims to generate sentences that describe the video automatically. It includes but is not limited to the tasks such as video captioning and video paragraph generation. Video captioning first dealt with SVO (Subject, Object, Verb) tuples-based models [23, 11, 22]. These approaches leverage objects and actions in the video and fit them into pre-defined sentence templates. With the success of deep learning models, later approaches [44, 49] frame the captioning task as one of the machine translation variants. In detail, Convolutional Neural Networks (CNNs) [25, 5, 46] are used as the encoders to model the visual data, and Recurrent Neural Networks (RNNs) are used as the decoders to generate the sentences. Chen *et al.* [7] use attention strategies to aggregate the entire video temporally to capture motion dynamics better. To alleviate the computational challenge of using expensive 3D CNNs applied to dense frame inputs, Seo *et al.* [37] use a Transformer-based encoder applied to raw pixels, sampled at a coarse rate to capture long context. To give models more information to understand the videos, some works [19, 20] take not only visual input but also utilize multi-modal inputs such as audio and speech. Video paragraph generation is generally viewed as a harder task compared with video captioning. It requires the model to handle long-term dependencies in the video and summarize the entire video content by generating multiple sentences. Liu *et al.* [27] propose a framework by modeling this task as a text summarization problem. The model first generates several sentence-level captions and then summarizes them into a paragraph. On the other hand, Song *et al.* [41] propose a one-stage framework for the video paragraph generation task by leveraging dynamic video memory and directly generating a paragraph.

Text-Video Retrieval. Text-Video retrieval aims to retrieve the most similar video given a textual query (or vice-versa). Numerous approaches to this problem revolve around offline feature extraction [50, 8, 15, 13]. However current methods mainly train deep models in an end-to-end manner [26, 14, 28, 29, 45, 4]. More recently, multiple papers [14, 29, 4] have proposed to leverage the CLIP [34] model as a backbone for text-video retrieval due to the high-level semantics encoded, and related to language, contained in the CLIP model. Apart from CLIP-based approaches, TS2-Net [28] proposes a token selection module that dynamically adjusts the selection criteria across temporal and

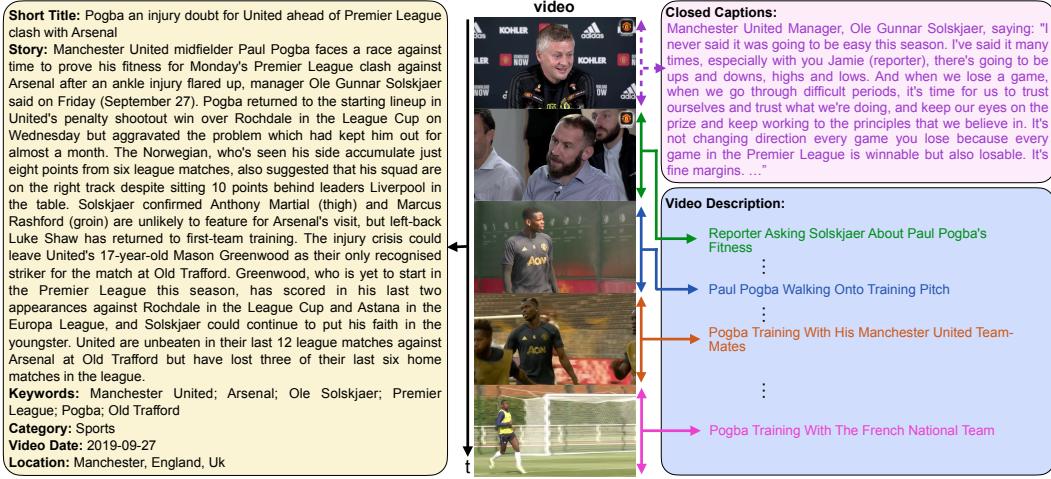


Figure 2: **Dataset details.** Here is an example from the ReutersViLNews dataset. ReutersViLNews contains the video and corresponding metadata, including short title, story, keywords, category, video description for video clips, closed caption, video date, and location. More details about the metadata information are in Section 3.1.

spatial dimensions of the input video, while DRL [45] exhaustively compares all input tokens and simultaneously learns to minimize the temporal redundancy of the sampled representation.

Video Language Datasets. The datasets for video-language tasks are the keys to the fast advancement of the research area. We compare the statistics and annotations of the existing datasets [54, 36, 40, 6, 48, 51, 24, 16, 30] in Table 1. The datasets are categorized into four main classes: Cooking, Movies, YouTube, and Social Media. In contrast, our proposed ReutersViLNews dataset focuses on the News domain which is different from existing video-language datasets. Most of these existing datasets are collected by scrapping the internet and are automatically labelled or labelled by crowdsourcing. Furthermore, they mostly contain single-sentence descriptions for one type of annotation, such as captions or stories. Ours, on the other hand, has a variety of annotations, such as captions, stories, keywords, etc. For example, the VideoStory dataset [16] includes multi-sentence descriptions of events visibly occurring in the video. Contrastingly, the story in ReutersViLNews contains the background information (e.g., location, date, and other stakeholders than the involved parties within the video) that cannot be directly inferred from the video frames and ultimately requires combining external knowledge with the available visual information.

3 Reuters Video-Language News Dataset

The Reuters Video-Language News Dataset (ReutersViLNews) is a dataset consisting of 1,974 long-form news videos with an average video length of 91.2 seconds. Each video reports on a single specific news story and may contain action shots of the event itself, interviews with people associated with the event, footage from nearby areas, and more. Figure 1 shows several examples from the dataset. The videos in ReutersViLNews are professionally produced, annotated, and fully licensed. The dataset has coverage from 200+ global locations and in 16 languages. Reuters is the world's largest international multimedia news agency and the leading provider of real-time news and intelligence. Moreover, a recent study by the Economist [2] showed that Reuters has the highest 'accuracy score' of all the publications included, and tracks at the center on 'bias score', showing neither left- nor right-wing bias, making suitable for both research and application purposes.

3.1 Labels and Metadata

Each video is in MP4 format and contains corresponding audio clips of reporting, interviews, etc. Each video is annotated by professionals and contains the following eight types of labels and metadata (examples are shown in Fig. 2). (1) **Short Title:** a one-sentence caption describing the news story covered in the video. (2) **Story:** A long multi-sentence description of the entire story in detail which

Table 2: **Dataset Statistics.** The first section is the video statistics. In order to keep the diversity of the dataset, we make sure each category has at least 200 videos. The second and third sections are the caption and story statistics, respectively. We show the average length of the text part and the percentage of the tokens covered by Glove-6B [33]. The final section is the dataset splits. We split the dataset to train/val/test with 70/15/15 ratio.

		Full	Entertain.	Misc	Disaster	Finance	Health	Politics	Sports
Stat.	#videos	1,974	200	224	200	325	346	479	200
	Avg. Length (s)	91.2	114.6	85.6	95.4	87.2	90.4	83.3	96.6
Story Cap.	Avg. Length	11.3	11.7	11.2	10.6	11.9	11.1	11.6	10.8
	% in Glove [33]	91.7	94.2	95.2	96.5	94.9	94.2	92.7	93.4
Story	Avg. Length	178.0	185.7	169.1	155.9	175.6	183.4	173.7	207.1
	% in Glove [33]	85.1	93.0	93.1	93.2	91.7	91.0	91.5	91.2
Splits	Train	1,382	142	153	134	227	253	342	131
	Validation	296	26	34	33	52	45	72	34
	Test	296	32	37	33	46	48	65	35

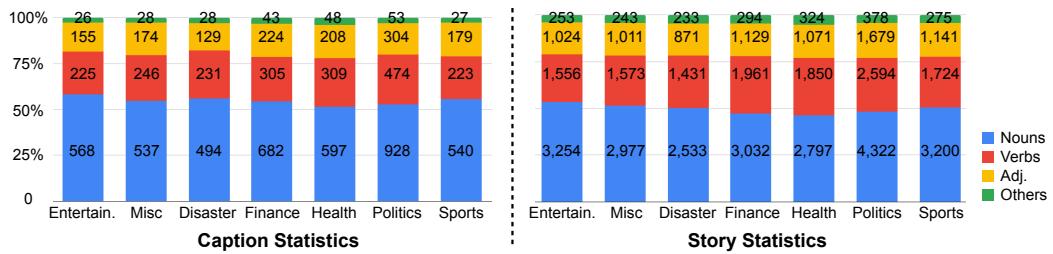


Figure 3: **Text Statistics.** We show the composition of captions and stories across different categories.

may contain contents that are not visually observable in the video, such as references to places, names, and events. (3) **Keywords:** Each video is labeled with multiple keywords used for identifying and tagging similar videos. (4) **News Category:** A manually labeled categorization (there are seven categories total) of the topic covered in the story (e.g., sports). (5) **Video Description:** Short phrases describing the visual phenomenon in the video grounded to each camera shot. (6) **Closed Captions:** Manually transcribed captions corresponding to speaking in the audio. (7) **Video Date.** (8) **Location.**

The ReutersViLNews dataset contains rich semantic labels and metadata for a diverse range of topics from around the world (see Sec. 3.3 for the quantitative statistics of ReutersViLNews). In this paper, our aim is to outline several challenging tasks on news dataset. However, we note that additional tasks will be possible with the dataset in its current state (e.g., as ground truth for text-to-video generation), or alternatively, if additional annotations are produced (e.g., temporal annotations for video grounding). A deeper discussion of these tasks and future directions are given in Secs. 4 and 5.

3.2 Labeling Guidelines

There are several differences in the labeling guidelines given to Reuters journalists (who have years or even decades of experience) and labeling strategies such as Mechanical Turk. First, journalists are given high-level guidelines when labeling videos, rather than fine-grained instructions on exactly what to label and what to ignore. The general instructions given to all journalists when labeling the videos are to prioritize the following factors in their labeling: objectivity, bias, truth, standards of integrity, the presentation of information, and details of how to maintain quality in specific situations or with specific kinds of reporting. Indeed, a complete set of rules that covers all situations that can occur on a daily basis is impossible to enumerate. It is worth re-emphasizing that Reuters journalists who create video content usually have years, and in some cases decades of experience providing consistent labels and annotations on video. Moreover, all Reuters journalists complete internal training to comply with those guidelines before serving in a news group under editorial staff. All published videos are also subject editorial review before publication to ensure compliance with the above guidelines. The videos are then distributed to thousands of news organizations around the planet. While inconsistencies or errors in labeling or metadata do sometimes occur in fast-breaking

Table 3: **Quantitative Results for Video Captioning.** The video captioning task is trained on two baselines, MDVC [20] and BMT [19], with two different settings, w/o and w/ audio input. We evaluate metrics with two traditional metrics, BLEU@n [32], METEOR [12].

	Model	BLEU3↑	BLEU4↑	METEOR↑
Visual-only	MDVC [20]	19.56	12.11	7.44
	BMT [19]	19.82	12.30	6.79
Visual + Audio	MDVC [20]	20.06	12.55	8.00
	BMT [19]	20.14	12.65	7.75



Figure 4: **Qualitative Results for Video Captioning.** We show two qualitative examples from each model and each setting. As shown in the figure, BMT [19] can generate more related captions compared with others. Anecdotally, one can also see that training with audio input describes the video more thoroughly, as ReutersViLNNews contains interviews and press conferences.

complex news narratives, Reuters dynamically iterates on the dataset labels, as this distribution and real world use results in any errors being reported back to us by our partners and fixed very quickly.

3.3 Dataset Statistics

This section provides a comprehensive description of the video dataset and its associated splits for training, validation, and testing. We adopt a balanced approach in selecting samples from each category to ensure that models trained on it can generalize well to unseen data. The training, validation, and testing splits are made up of 70% (1,382 videos), 15% (296 videos), and 15%, respectively. We select videos such that all categories contain a minimum of 200 videos in order to reduce the negative impacts of categorical bias.

Table 2 and Fig. 3 provide further details about the number of videos in each category and their average length. Each video in the dataset is annotated by a caption, a long title, and a set of keywords that can be used for more granular analysis. Furthermore, it is noteworthy that the distribution of the data is fairly uniform, which assists machine learning models and prevents data imbalances from skewing the results of the model.

4 Experiments

In this section, we provide benchmarks for a meaningful sampling of the video-language tasks possible using ReutersViLNNews. For this purpose, we present experimental results for four varied tasks on the dataset: (i) video captioning (Sec. 4.1), (ii) video paragraph generation (Sec. 4.2), (iii) open-world keyword generation (Sec. 4.3, and (iv) text-video retrieval (Sec. 4.4). Note that we specifically chose tasks such that they vary in their difficulty (e.g., video captioning vs. paragraph generation), output (e.g., generative vs. discriminative), and input (e.g., text vs. video vs. audio).

4.1 Task 1: Video Captioning

Given a video, the video captioning task aims to generate a one-sentence description of the video. We train and evaluate two baselines, BMT [19] and MDVC [20] on ReutersViLNNews. We take the **Short Title** as our ground truth caption (see Fig. 2 for an example). For both models, we follow the same training settings as in the original papers except we change the dictionary size to 3,841 and use the ground truth proposals. We follow the original papers and train the models with two different settings: (i) visual-only and (ii) visual+audio.

The results for both models and settings are shown in Table 3 where the performance of both models is evaluated using BLEU@n [32] score and METEOR [12] score. We also visualize the results from

Table 4: **Quantitative Results for Video Paragraph Generation.** We train the state-of-the-art video paragraph generation model [41] on ReutersViLNews and evaluate on three domains, accuracy, diversity, and similarity. $B@4$, M , P_B , R_B , and F_B are BLEU@4, METEOR, P_{BERT} , R_{BERT} , and F_{BERT} respectively. FT denotes the model is finetuned on ReutersViLNews.

Model	Accuracy		Diversity			Similarity	
	$B@4 \uparrow$	$M \uparrow$	$Div@1 \uparrow$	$Div@2 \uparrow$	$Rep@4 \downarrow$	$Sen-Sim \uparrow$	$F_B \uparrow$
Song <i>et al.</i> [41]	0.01	2.7	69.7	85.5	2.2	4.9	80.1
Song <i>et al.</i> [41] (FT)	0.7	5.4	73.8	87.1	2.1	33.5	81.1

different settings in Fig. 4. The results show that using visual and audio data performs better than only using visual data only. It is expected as ReutersViLNews contains many clips of audio-centric reporting events, such as press conferences and interviews.

4.2 Task 2: Video Paragraph Generation

Given a video, the goal of the video paragraph generation task is to generate a multi-sentence paragraph that plausibly describes the video. Compared to the video captioning task (Sec. 4.1), this task produces longer descriptions that offer a more detailed description over longer time periods. In this task, we train and evaluate the state-of-the-art video paragraph generation model by Song et al. [41] on ReutersViLNews. The input of the model is the entire video, and we take the **Story** (see Fig. 2 for an example) as the target of the generated paragraph. A key component of the model is the keyframe selection module, which learns to subsample important frames from the video to reduce computation. We follow the same training and evaluation settings as in the original paper.

The paragraph generation results are shown in Table 4. We first evaluate the performance using accuracy (BLEU@n and METEOR score), and diversity (n-gram diversity [39]: $Div@n$, and n-gram repetition [47]: $Rep@n$) metrics. Interestingly, while the diversity is strong, the accuracy (i.e., BLEU@4 and METEOR score) for both models is quite low. Although finetuning increases the accuracy slightly, the scores are not high enough to draw any meaningful conclusions, e.g., the BLEU@4 is only 0.7. We hypothesize that, due to the abstract and high-level nature of this dataset, these accuracy metrics are incapable of measuring adequate *semantic* similarity between the prediction and target paragraphs. Towards a solution to this problem, we evaluate the models based on semantic similarity metrics (Sentence-Similarity [35] and Bert F-score [53] (F_{BERT})). These metrics are based on large language model encodings of the prediction and target paragraphs. As expected, these model-based similarity metrics demonstrate that finetuning improves the model output significantly. These results agree with the qualitative results observed in Fig. 5, where the finetuned model makes less mistakes and captures more appropriate context than the non-finetuned model.

4.3 Task 3: Open World Keyword Generation

Given a video, the task of keyword generation is to tag the video with keywords that describe the content of the video. This is a difficult task because (i) the task has an open vocabulary and (ii) there is a potentially unbounded number of keywords (however, in practice, the upper limit of keywords in the dataset is around seven). We implement two baselines for this task. For the first baseline, we treat the task as a supervised classification problem and set the number of output logits of a classification model to the total number of keywords. However, note that this baseline is bounded above due to the fact that the validation and test sets contain keywords that are not found in the training set. For the model architectures, we modify 2D ResNets [18] to take 12 uniformly sampled frames from the video and then apply spatio-temporal global average pooling to the features before the linear layer. The second baseline is to use the CLIP model [34] in a zero-shot setting. Following the original paper, we query the CLIP model with the phrase “*a photo of {}/*” for each frame in a video, and average the score across all frames. We then take all keywords whose score sits above the 95th percentile of the maximum logit value.

The keyword generation results are presented in Table 5. As expected, the resulting F1 scores are quite low for all baselines. Interestingly, the supervised results outperform the zero-shot baselines, even though the performance of the supervised setting is bounded above based on the intersection of



Ground Truth: firefighters in the australian state of tasmania battled against 27 bushfires across the state on wednesday january 23. tasmania fire service deputy chief officer, bruce byatt, told reporters that fire personnel were facing 720 kilometers of firefront in the state, with 55,000 hectares of bushland and one property already lost. firefighting aircraft were sent in from the state of new south wales to help in the fight. australia is currently in the midst of the southern hemisphere summer and experiencing heatwave conditions stifling southeastern parts of the country tasmanians are bracing for temperatures to rise again on friday january 25.

Song et al.: a camera pans around a large yard and leads into a man speaking and people getting ready to fly . several shots are shown of people smoking and waving to the camera as well as people celebrating and speaking to the camera .

Song et al. (FT): australian police said on thursday february 10 that two people have been destroyed in two people in bushfires in the east of perth, as authorities look to contain the western australia. several emergency warnings that have been destroyed in place since the fire season and queensland fire service six homes were destroyed in place in queensland state.

Figure 5: **Qualitative Results for Video Paragraph Generation.** We visualize the video and output paragraph from the baseline [41] and highlight the related content in green.

Table 5: **Open World Keyword Generation.** We evaluate zero-shot and supervised baselines for the task of open-world keyword generation on ReutersViLNews. This task poses significant challenges given the rare occurrences of many keywords.

Model	Zero-Shot	Recall	Precision	F1
ResNet18 [18]	✗	0.081	0.493	0.137
ResNet50 [18]	✗	0.078	0.370	0.124
CLIP R50 [34]	✓	0.082	0.078	0.076
CLIP ViT-b [34]	✓	0.092	0.073	0.079

the keywords between the training and validation set. We observe that ResNet18 slightly performs ResNet50 and that both CLIP baselines perform comparably to each other. The results suggest that these simple baselines are not sufficient for solving the task of open-vocabulary keyword generation.

4.4 Task 4: Video-Text Retrieval

Given a query video, the task of Video-Text Retrieval is to return the most closely associated textual description (i.e., caption) corresponding to the query (and vice-versa for Text-Video retrieval). Example applications of this task are retrieval search engines in large unlabelled video databases. For this task, we train three recent baselines: Clip4Clip [29], DRL [45], and TS2-Net [28]. While the captions in traditional datasets (e.g., MSR-VTT [48], ActivityNet [24]) focus on describing actions occurring in the video, ReutersViLNews captions contain higher-level descriptions of the events and story contained in the video (see Fig. 2 for examples of the Short Title). All models are trained with 12 frames evenly spaced as input from random spatial and temporal cropped video segments.

The results for the Text-to-Video (T2V) and Video-to-Text (V2T) retrieval tasks are presented in Table 6. The ReutersViLNews dataset is a challenging dataset compared to traditional T2V retrieval benchmarks of similar size. For example, despite MSR-VTT [48] containing almost twice as many validation videos as ReutersViLNews (1,000 vs. 632), TS2-Net achieves an R@1 of 47.0% on MSR-VTT compared with 51.4% on ReutersViLNews. Given the smaller number of examples to choose from during inference, we conclude that ReutersViLNews poses a difficult challenge for the text-video retrieval task. Figure 6 visualizes the interpretable text-video attention mechanism from the DRL [45] model. Interestingly, the model learns to recognize and attend to the UK Prime Minister, Theresa May, speaking in parliament in order to rank this video as likely belonging to the caption. This exemplifies how high-level visual phenomena, such as specific people (e.g., country leaders) and places (e.g., specific parliamentary buildings) are important for solving the ReutersViLNews dataset.

Table 6: **Quantitative Results for Text-Video Retrieval.** We train three state-of-the-art text-video retrieval models on ReutersViLNews. The model performances reveal that the dataset is challenging to solve and TS2-Net [28] outperforms the other models.

	Model	R@1	R@5	R@10	MnR
Text-to-Video Retrieval	DRL [45]	36.2	60.4	72	15.0
	Clip4Clip [29]	42.6	72.3	82.4	7.5
	TS2-Net [28]	51.4	80.4	90.5	5.5
Video-to-Text Retrieval	DRL [45]	30.9	59	71.4	14.8
	Clip4Clip [29]	39.5	72.0	86.1	6.3
	TS2-Net [28]	52	80.7	89.5	5.2

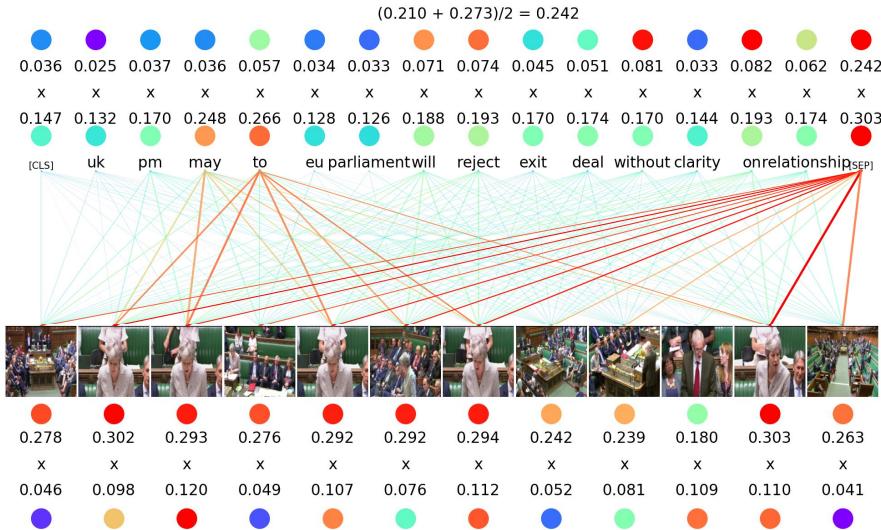


Figure 6: **Qualitative Results for Text-Video Retrieval.** We visualize the text-video token attention maps from the DRL model [45] for a single video. Notice how the model attends to Theresa May over multiple frames, suggesting the model learns to recognize specific people who appear many times in the dataset.

5 Conclusion

This technical report presents an analysis on a large-scale video-language understanding dataset, the Reuters Video-Language News (ReutersViLNews) dataset. ReutersViLNews consists of 1,974 news-oriented videos covering seven diverse categories and are collected and labeled by professionals in the news industry. Each video contains a video caption (short title), a story that describes the content of the video, keywords, a video category, detailed video descriptions, closed captions, video date, and location. We benchmark the ReutersViLNews dataset on four different video-language tasks: (i) video captioning, (ii) video paragraph generation, (iii) open-world video keywords generation, and (iv) video-text retrieval. Experiments were run for each of these tasks and revealed several interesting findings and directions for future work. For the video-captioning task (Sec. 4.1), the audio modality can improve performance however it is likely that this can be further improved with better mixing of audio-visual information. For video paragraph generation (Sec. 4.2), the BLEU and Meteor metrics are insufficient for measuring ‘story’ similarity. More agreeable metrics are Sentence-Similarity [35] and Bert F-score [53], which, when used, show the benefits of fine-tuning on ReutersViLNews. Keyword generation was shown to be a very difficult task for *both* supervised and open-world vision language models. Finally, visualizations showed that text-video retrieval models leverage identifying specific people who appear frequently in the dataset in order to associate the corresponding caption (or video).

Acknowledgement

The Reuters News Agency and Thomson Reuters members and subject matter experts deserve special credit. We thank Yulia Pavlova for managing the delivery of the dataset and for her invaluable feedback and support throughout this project. We would like to thank Vector Institute for making this collaboration possible and providing academic infrastructure and computing support during all phases of this work.

References

- [1] Sami Abu-El-Haija et al. “Youtube-8m: A large-scale video classification benchmark”. In: *ArXiv* (2016).
- [2] Stephen J Adler. “Reuters features in Economist study on accuracy and bias”. In: (2019).
- [3] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *CVPR*. 2018.
- [4] Max Bain et al. “A CLIP-Hitchhiker’s Guide to Long Video Retrieval”. In: *ArXiv* (2022).
- [5] Joao Carreira and Andrew Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *CVPR*. 2017.
- [6] David Chen and William B Dolan. “Collecting highly parallel data for paraphrase evaluation”. In: *ACL: human language technologies*. 2011.
- [7] Shaoxiang Chen and Yu-Gang Jiang. “Motion guided spatial attention for video captioning”. In: *AAAI*. 2019.
- [8] Shizhe Chen et al. “Fine-grained video-text retrieval with hierarchical graph reasoning”. In: *CVPR*. 2020.
- [9] Shih-Han Chou, James J Little, and Leonid Sigal. “Implicit and Explicit Commonsense for Multi-sentence Video Captioning”. In: *ArXiv* (2023).
- [10] Shih-Han Chou et al. “Semi-supervised Grounding Alignment for Multi-modal Feature Learning”. In: *CRV*. 2022.
- [11] Pradipto Das et al. “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching”. In: *CVPR*. 2013.
- [12] Michael Denkowski and Alon Lavie. “Meteor universal: Language specific translation evaluation for any target language”. In: *Proceedings of the ninth workshop on statistical machine translation*. 2014.
- [13] Maksim Dzabraev et al. “Mdmmmt: Multidomain multimodal transformer for video retrieval”. In: *CVPR*. 2021.
- [14] Han Fang et al. “Clip2video: Mastering video-text retrieval via image clip”. In: *ArXiv* (2021).
- [15] Valentin Gabeur et al. “Multi-modal transformer for video retrieval”. In: *ECCV*. 2020.
- [16] Spandana Gella, Mike Lewis, and Marcus Rohrbach. “A dataset for telling the stories of social media videos”. In: *EMNLP*. 2018.
- [17] Tengda Han, Weidi Xie, and Andrew Zisserman. “Temporal alignment networks for long-term video”. In: *CVPR*. 2022.
- [18] Kaiming He et al. “Deep residual learning for image recognition”. In: *CVPR*. 2016.
- [19] Vladimir Iashin and Esa Rahtu. “A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer”. In: *BMVC*. 2020.
- [20] Vladimir Iashin and Esa Rahtu. “Multi-Modal Dense Video Captioning”. In: *CVPR Workshops*. 2020.
- [21] Md Mohaiminul Islam and Gedas Bertasius. “Long movie clip classification with state-space video models”. In: *ECCV*. 2022.
- [22] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. “Natural language description of human activities from video images based on concept hierarchy of actions”. In: *IJCV* (2002).
- [23] Dieter Koller, Norbert Heinze, and Hans-Hellmut Nagel. “Algorithmic characterization of vehicle trajectories from image sequences by motion verbs.” In: *CVPR*. 1991.
- [24] Ranjay Krishna et al. “Dense-captioning events in videos”. In: *ICCV*. 2017.
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *ACM* (2017).

- [26] Jie Lei et al. “Less is more: Clipbert for video-and-language learning via sparse sampling”. In: *CVPR*. 2021.
- [27] Hui Liu and Xiaojun Wan. “Video paragraph captioning as a text summarization task”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2021.
- [28] Yuqi Liu et al. “TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval”. In: *ECCV*. 2022.
- [29] Huaishao Luo et al. “CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning”. In: *Neurocomputing* (2022).
- [30] Antoine Miech et al. “HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips”. In: *ICCV*. 2019.
- [31] Mathew Monfort et al. “Moments in time dataset: one million videos for event understanding”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.2 (2019), pp. 502–508.
- [32] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *ACL*. 2002.
- [33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In: *EMNLP*. 2014. URL: <http://www.aclweb.org/anthology/D14-1162>.
- [34] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *ICML*. 2021.
- [35] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *EMNLP*. 2019.
- [36] Anna Rohrbach et al. “A dataset for movie description”. In: *CVPR*. 2015.
- [37] Paul Hongseok Seo et al. “End-to-end generative pretraining for multimodal video captioning”. In: *CVPR*. 2022.
- [38] Piyush Sharma et al. “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning”. In: *ACL (Volume 1: Long Papers)*. 2018.
- [39] Rakshith Shetty et al. “Speaking the same language: Matching machine to human captions by adversarial training”. In: *ICCV*. 2017.
- [40] Gunnar A Sigurdsson et al. “Hollywood in homes: Crowdsourcing data collection for activity understanding”. In: *ECCV*. Springer. 2016.
- [41] Yuqing Song, Shizhe Chen, and Qin Jin. “Towards Diverse Paragraph Captioning for Untrimmed Videos”. In: *CVPR*. 2021.
- [42] Zhan Tong et al. “VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training”. In: *NeurIPS*. 2022.
- [43] Tomoki Uchiyama et al. “Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis”. In: *WACV*. 2023.
- [44] Subhashini Venugopalan et al. “Sequence to sequence-video to text”. In: *ICCV*. 2015.
- [45] Qiang Wang et al. “Disentangled Representation Learning for Text-Video Retrieval”. In: *ArXiv* (2022).
- [46] Saining Xie et al. “Rethinking spatiotemporal feature learning for video understanding”. In: *ArXiv* (2017).
- [47] Yilei Xiong, Bo Dai, and Dahua Lin. “Move forward and tell: A progressive generator of video descriptions”. In: *ECCV*. 2018.
- [48] Jun Xu et al. “Msr-vtt: A large video description dataset for bridging video and language”. In: *CVPR*. 2016.
- [49] Li Yao et al. “Describing videos by exploiting temporal structure”. In: *ICCV*. 2015.
- [50] Youngjae Yu, Jongseok Kim, and Gunhee Kim. “A joint sequence fusion model for video question answering and retrieval”. In: *ECCV*. 2018.
- [51] Kuo-Hao Zeng et al. “Title generation for user generated videos”. In: *ECCV*. Springer. 2016.
- [52] Chen-Lin Zhang, Jianxin Wu, and Yin Li. “ActionFormer: Localizing Moments of Actions with Transformers”. In: *ECCV*. 2022.
- [53] Tianyi Zhang* et al. “BERTScore: Evaluating Text Generation with BERT”. In: *ICLR*. 2020.

- [54] Luowei Zhou, Chenliang Xu, and Jason Corso. “Towards automatic learning of procedures from web instructional videos”. In: *AAAI*. 2018.