

---

# MultiVENT: Multilingual Videos of Events with Aligned Natural Text

---

Kate Sanders\*   David Etter\*   Reno Kriz\*   Benjamin Van Durme

Johns Hopkins University  
Human Language Technology Center of Excellence  
{ksande25, detter2, rkriz1, vandurme}@jhu.edu

## Abstract

Everyday news coverage has shifted from traditional broadcasts towards a wide range of presentation formats such as first-hand, unedited video footage. Datasets that reflect the diverse array of multimodal, multilingual news sources available online could be used to teach models to benefit from this shift, but existing news video datasets focus on traditional news broadcasts produced for English-speaking audiences. We address this limitation by constructing MultiVENT, a dataset of multilingual, event-centric videos grounded in text documents across five target languages. MultiVENT includes both news broadcast videos and non-professional event footage, which we use to analyze the state of online news videos and how they can be leveraged to build robust, factually accurate models. Finally, we provide a model for complex, multilingual video retrieval to serve as a baseline for information retrieval using MultiVENT.

## 1 Introduction

Information dissemination for current events has traditionally consisted of professionally collected and produced materials, leading to large collections of well-written news articles and high-quality videos. As a result, such materials form the basis for significant prior work in content analysis and retrieval [52, 20, 2, 15, 48]. Meanwhile, a high volume of event-centric content today is generated by non-professionals, such as on-the-scene witnesses to events who hastily capture videos and upload them to the internet without further editing. We propose that this contemporary landscape of news content can be leveraged by models to produce a more comprehensive understanding of events. News agencies have adapted to this shift, often collecting and incorporating this online content into official broadcasts, but news video datasets still do not typically address this new domain of event coverage.

In addition to focusing on traditional news sources, existing news video datasets predominantly consider content produced in English. This is consistent with common practices in video dataset collection: Collected videos and captions are recorded in English, and when multilinguality is considered, it is achieved by directly translating captions and transcripts [46, 23, 38, 19]. Because this data is originally produced for English speaking audiences, these multilingual datasets can contain unwanted content biases like "translationese" [6, 21]. As event-centric video content produced in other languages makes up a large portion of news videos online, we argue that including organic, multilingual content is necessary for a diverse and perspective-agnostic sampling of event coverage.

With these ideas in mind, we present MultiVENT, a dataset of **Multilingual Videos of Events** with aligned **Natural Text** that contains 2,396 diverse, event-centric videos and text descriptions

---

\*Equal contribution.

## Query

بعد جولات من الحوار رعتها بغداد وسلطنة عمان لتكلمها الصين بفتح باب جديد. الاتفاق بين السعودية وإيران على إعادة العلاقات الدبلوماسية وفتح السفارات في البلدين

**Translation:** After rounds of dialogue sponsored by Baghdad and the Sultanate of Oman, China crowned it with the opening of a new door... The agreement between Saudi Arabia and Iran to restore diplomatic relations and open embassies in the two countries.



Figure 1: A sample video-text pair from MultiVENT. Every event-centric video is paired with a corresponding video description and a long-form text document describing the event, both in the same language as the video. If the language is not English, the video is also paired with a corresponding English document.

that reflect the distribution of news content online. The videos are grounded in natural language video descriptions and long-form text documents, and the data spans 260 current events across over forty countries. The content in MultiVENT is collected in five target languages: Arabic, Chinese, English, Korean, and Russian, and as the multilinguality is organic, the data is less likely to suffer from translation bias. We provide an illustration of the dataset’s contents in Figure 1: Each natural language query (describing a video of a current event) is paired with grounding text documents and a unique corresponding video. We use MultiVENT to explore and characterize the variety of event-centric videos available online and illustrate the importance of leveraging these different video types when building multimodal information systems.

Citizen journalism, the most notable example being Wikipedia [18], has emerged alongside other online news sources as a method for curating comprehensive summaries of events. Work in natural language processing has considered the problem of automating this process by training models to generate informative reports using online source materials [25, 43, 32]. We use MultiVENT to explore how this process can be extended to incorporate multimodal sources of evidence. As a first step in this direction, we consider the task of video retrieval on MultiVENT, through which a model learns to retrieve multimodal source material given a natural language event description. This task differs from prior video retrieval benchmarks [9, 51, 1, 28, 49] as the videos in MultiVENT vary widely in length and content presentation, are multilingual, and can involve significant amounts of on-screen text. In addition to multilingual natural language captions for each video, we provide full text documents that ground the events and serve as more complex retrieval queries.

In summary, our contributions are:

1. We present MultiVENT, a multimodal, multilingual information retrieval dataset of grounded videos depicting current events. The dataset targets five languages and covers a range of online video formats beyond traditional news broadcasts.
2. Using MultiVENT, we quantitatively illustrate the information presented by news videos and the differences in content between video formats, and qualitatively evaluate how multimodal coverage of an event can evolve over time.
3. We present MultiCLIP, a model for multilingual, event-centric video retrieval that serves as a baseline for video retrieval approaches on the task.

## 2 Related Work

### 2.1 Video retrieval datasets

Early video datasets generally contained short clips spanning narrow ranges of topics, such as the Microsoft Research Video Description Corpus [9]. Video datasets spanning larger domains include

MSR-VTT [51] and DiDeMo [1], although the lengths of these videos were still relatively short. The V3C dataset [37, 3] offered longer video lengths while still spanning a wide range of topics such as news reports. A shift towards massive video datasets was instigated by HowTo100M [28], which included over 130 million video clips belonging to one million narrated instructional videos. VaTeX [46], released in the same year, considered video retrieval from a multilingual context using caption translation. Additional multilingual video retrieval datasets include Rudder [13], consisting of instructional videos for making toys with multilingual captions, MTRV [23], which extended the TVR dataset [24] by adding Chinese subtitles and queries, and Multi-HowTo100M [19], which extended HowTo100M by scraping YouTube for subtitles in up to 9 other languages. Recently, Chen et al. [8] released the ChinaOpen dataset which contains a wide range of video-caption pairs originally produced in Chinese. Recent work has also considered the problem of interpreting text-heavy video content: Wu et al. [49] and Jahagirdar et al. [20] introduced datasets that focus on within-video text and OCR annotations, including news broadcasts.

## 2.2 Video retrieval methods

The size of early video datasets allowed retrieval systems to rely on pre-extracted features from expert systems like action recognition models. As massive video datasets gained prominence, the video retrieval paradigm moved towards ad-hoc video-text feature extraction using large pretrained models. Dosovitskiy et al. [14] proposed using stand-alone transformer architectures for video understanding, and Bertasius et al. [7] showed that applying space- and time-based self-attention independently improved performance. Bain et al. applied findings directly to video retrieval, training and evaluating transformer architectures on WebVid-2M [4]. Radford et al. [33] introduced CLIP and showed that pretraining models to match captions to images can result in scalable models, and CLIP’s applicability to video retrieval was demonstrated by Fang et al. [16] through their CLIP2Video model. More fine-grained modifications to CLIP were proposed. Wang et al. [45] introduced "Object-aware Transformers", which extended video-text transformers to incorporate object-level annotations within video footage, and Ge et al. [17] modified the pretraining task to involve teaching a vision-text model to answer multiple choice questions about a video. Bain et al. [5] adapted large image-text models to the task of long video retrieval by incorporating the weighted-mean of frame embeddings, and Wu et al. [49] incorporated independent optical character recognition and embeddings into the encoder pipeline to explicitly model in-video text.

## 2.3 Report generation using online sources

A wide range of research has used online corpora for report generation tasks, including QA-pair and knowledge graph generation [35, 34, 53, 29, 22, 30]. Notably, Lewis et al. [25] introduced a method for automatically extracting question-answer pairs from large corpora of text documents, and applied this method to Wikipedia to produce the PAQ dataset. Some PAQ extensions have been multilingual — Pisare et al. [31] built the WikiOmnia QA dataset on Russian Wikipedia documents, and Rybak et al. [40] produced a question-Wikipedia passage dataset in Polish. Recently, Qian et al. [32] extended the ideas in PAQ to construct WebBrain, a task in which a model must generate factual articles with references given a natural language query. In the multimodal domain, Reddy et al. and Chen et al. have considered the problem of open-domain QA for image-text data [36, 10], with Chen et al. using Wikipedia to generate a multimodal dataset. In a similar vein, Li et al. propose a dataset for information extraction from multimedia articles [27] and an extraction approach that can be used with text, image, and video content [26].

## 3 Dataset

In this section we outline the MultiVENT collection process. The dataset includes 2,396 videos and corresponding text descriptions covering 260 current events grounded in 468 text documents, and includes content in Arabic, Chinese, English, Korean, and Russian. We first identify 260 visually salient current events spanning from 2013 to 2023, and assign a target language to each event. Then, for each event, we collect grounding text documents and a set of videos in the event’s target language.

### 3.1 Current event curation

We consider four primary event categories for MultiVENT: Disasters, political events, social events, and technology events. We include thirteen current events per category for each target language. We use Google Trends statistics to select these events, based on its tracking of term popularity based on internet activity by country. We construct lists of the top five countries with the most speakers of each target language and review the top trending topics on Google in each of these countries over the last ten years. We record topics and search phrases that corresponded to current events that (1) align with one of the predefined event categories and (2) have sufficient online video coverage. For categories that did not amass a sufficient list of current events per language through this process, we consult Wikipedia’s yearly summaries of events to fill the remaining slots. Detailed statistics characterizing this set of current events are shown in Figure 2. As shown, the majority of selected events take place in the last few years, with only three taking place before 2016.

Also shown in Figure 2, there is not a bijective mapping between the language used in event coverage and the country the event took place in. The language and country are often related, e.g., Russian news content in MultiVENT predominantly takes place in Russia, but this is not true of all events in the dataset. For example, we include data in Chinese pertaining to the 2023 ATP tennis circuit in Dallas, Texas: At this event, tennis player Wu Yibing became the highest-ranked Chinese player in the history of the ATP rankings, and so the event received substantial Chinese news coverage. In cases such as this, news in multiple languages will heavily focus on the same current event, such as sports events and international political relations. We do not include the same event in multiple languages in MultiVENT by design, in contrast with data collection procedures used for efforts such as AIDA [44] which aim to cover a small collection of current events in many languages.

Every current event in the dataset is grounded in an English natural language document and, if the event is tagged with a non-English language, an additional natural language document in that target language. First, we check if a full English Wikipedia article exists for the current event. If not, we manually find a Wikipedia article that includes a passage describing the event. If Wikipedia does not have a passage that appropriately grounds the event, then a news article in English is selected as a grounding document instead. This process is then repeated for the target language. The dataset includes 468 grounding articles in total: 313 are full Wikipedia articles, 104 are Wikipedia passages, and 51 are external articles.

### 3.2 Video collection

We aim to collect visually and semantically distinct videos for each current event with an even split between firsthand witness accounts (e.g., first-person smartphone videos), amateur edited videos (e.g., vlogs), and professional news reports and compilations. Information regarding the resultant distribution of these categories and their semantic differences is included in Section 4.2. For each current event, we collect ten videos in the current event’s target language. We search YouTube and Twitter for these videos using target keywords collected from the Google Trends search and Wikipedia. After collecting the videos, we manually identify and remove duplicates, resulting in 2,396 videos in total. We do not include repeat videos, but sometimes professional news reports include firsthand footage that is already included as unedited footage in the dataset. In these cases, we keep both the news report and the original footage as the context and text metadata between the two are distinct. If the video has a natural language description, we tag the video with this description. If it does not, we use the video title as the tagged natural language description. We report the distribution of videos by source in Figure 2.

## 4 Data analysis

We present an analysis of MultiVENT to help characterize how online multimodal content contributes to our understanding of current events. We explore multimodal event coverage from three angles: (1) what kinds of information news videos contribute, (2) the differences in informative content provided by different types of news videos, and (3) how multimodal coverage of an event can evolve over time.

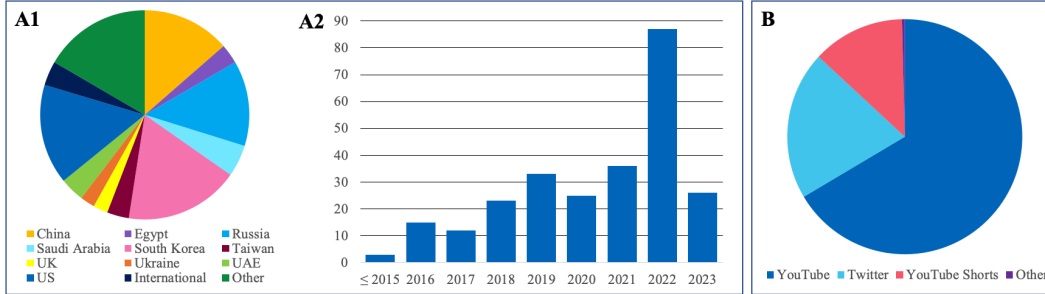


Figure 2: **A**: Statistics illustrating the distribution of current events selected for the dataset. (**A1**) depicts the general breakdown of countries in which each current event takes place. Many countries had a single current event, particularly small countries in the middle east and southeast Asia, and are consolidated into the "other" category for easier graph interpretation. (**A2**) shows the distribution of years during which the current events take place. **B**: Breakdown of data sources for the videos in the dataset. The majority of videos came from YouTube as it has a larger international audience and has existed longer than YouTube Shorts.

#### 4.1 Semantic information in video

Visual data can provide rich, semantically nuanced details of an event that are not captured in text documents due to reporting bias, limitations of text, and document length limits. To characterize the complexity of these videos and the information they provide, we annotate a set of two hundred videos of disasters in MultiVENT to identify visual entities in the videos that help answer common "who, what, where"-type questions about the events they depict.

We present videos of disaster footage to local annotators and provide them with a set of event-centric questions derived from FrameNet’s "disaster scenario" template [39]. We modify this template, designed to annotate the event semantics of text documents, to better cover the range of information provided by visual content. We instruct annotators to identify every on-screen entity (such as people, scrolling news headline banners, etc.) that might help answer one of these event-centric questions.

The template divides salient entities into six categories: The disaster itself ("what"), the location of the disaster ("where"), the time the disaster takes place ("when"), people affected by the disaster ("who") and first responders for the disaster, e.g., firefighters (also "who"), and any visible outcomes of the disaster. Not every category applies to both visual content and text: We exclude "where" and "when" from the set of categories that visual content should be annotated for (because identifiable depictions of "where" are present in almost every frame, and "when" in virtually none) and disaster outcomes from the set of text annotation categories, as textual examples of this category tend to involve full clauses, which complicate the annotation process.

We present the number of event-relevant entities that appear on-screen in these annotated videos in Table 1. For each annotated entity, we additionally ask annotators to rate their certainty that the entity is directly related to the event described by the video’s natural language description from 0% to 100%. We record these certainty scores in 20% intervals, i.e. as 20%, 40%, 60%, 80%, or 100%. The averages of the linguists’ confidence rankings by entity type are listed in Table 2.

As shown in Table 1, each video contains an average of 9.32 informative visual entities that pertain to the event in question. About half of these entities are purely visual, and half are within-video text that can be identified with an optical character recognition model. As indicated by Table 2, purely visual entities are more ambiguous than the text content shown onscreen alongside it, which aligns with past research that explores the difficulty humans have in interpreting visual content depicting complex events [41].

#### 4.2 Video content by domain

As described in Section 3, we collect three main types of videos: Official news broadcasts, edited video footage, and raw, unedited footage. Of the 210 videos in the annotation set reported in Table 1, 53% are news broadcasts, 11% are edited footage, and 36% are raw footage. To quantify the

Table 1: Mean number of visual entities and in-scene text references (written text displayed within a video) present per video in a subset of 210 disaster videos from the current events dataset. We omit "where" and "when" entities from the visual content counts as "where" visual content technically appears in every frame and there are few types of visual evidence for "when" questions. We omit "outcomes" from the text references as an outcome by itself is a full event that is difficult to localize in text (this field is omitted from the FrameNet event template analogue for text documents).

	Visual entities	Text references	Total
Disaster ("What")	1.25	1.37	<b>2.62</b>
Place of occurrence ("Where")	-	1.54	<b>1.54</b>
Time of occurrence ("When")	-	0.77	<b>0.77</b>
Affected people ("Who")	1.22	0.54	<b>1.76</b>
First responders ("Who")	1.13	0.50	<b>1.63</b>
Disaster outcomes	1.00	-	<b>1.00</b>
Total	<b>4.60</b>	<b>4.72</b>	<b>9.32</b>

Table 2: Mean annotator certainty scores partitioned on entity type based on the annotations used for Table 1. 0.20 certainty indicates that the annotator is 20% sure that the annotated entity helps answer the tagged question about the described event, while 1.00 certainty indicates that the annotator is completely sure that the entity helps answer the tagged question about the event.

	Disaster	Where	When	AP	FR	Outcomes	All
Visual content	.787	-	-	.716	.765	.798	.830
Text content	.931	.907	.929	.856	.836	-	.900
Average	.862	.907	.929	.759	.787	.798	.865

difference in information presented by these different video types, we take the video annotations shown in Table 1 and partition these annotations by video type. We present the results in Table 3.

As shown by the results, news broadcasts depict the most relevant semantic information, followed by edited footage. This is particularly apparent when considering text content alone. On average, news coverage contains almost 9 times as much relevant on-screen text content than raw footage, and over three times more than edited footage. Visual content differences were less drastic, but news content still had two times more visual content than raw footage and 1.3 times more than edited footage. The difference in visual content between news coverage and edited footage is possibly due to average video length and the quality of the video curation — oftentimes, unprofessionally edited footage only draws from one source whereas news coverage draws from many.

### 4.3 Information evolution

As shown in Table 3, first-person footage is often opaque compared to professional coverage. However, comprehensive coverage often builds on earlier, less informative coverage. This can be seen in news cycles for slowly unfolding events and for sudden, unexpected events that take time to assess. This is illustrated in Figure 3, which shows a snapshot of the 2019 Notre Dame fire news cycle and demonstrates how unedited and poorly curated footage, often first-person witness accounts on social media, can be instrumental in the construction of our collective understanding of events. So, we propose that teaching models to understand different video formats, despite clear discrepancies in the amount of information they present, is important for developing robust systems.

## 5 Experiments

### 5.1 Approach

We consider the problem of teaching a model to map multilingual, natural language queries to multilingual video clips. Specifically, we consider a video set  $V$  and query set  $T$  with an indicator mapping function  $f$  that returns whether a query  $t \in T$  describes a video  $v \in V$ . The model  $h$  is

Table 3: Mean number of visual entities and in-scene text references present per video, partitioned on video type. Same 210 video subset is used for analysis as that used for the analysis shown in Table 1.

	News coverage			Edited footage			Raw footage		
	Vis.	Text	Total	Vis.	Text	Total	Vis.	Text	Total
Disaster ("What")	1.42	2.38	<b>3.80</b>	1.14	0.41	<b>1.55</b>	1.05	0.17	<b>1.22</b>
Place ("Where")	-	2.51	<b>2.51</b>	-	0.59	<b>0.59</b>	-	0.39	<b>0.39</b>
Time ("When")	-	1.26	<b>1.26</b>	-	0.41	<b>0.41</b>	-	0.16	<b>0.16</b>
Affected people ("Who")	1.48	0.94	<b>2.42</b>	1.18	0.23	<b>1.41</b>	0.86	0.04	<b>0.90</b>
First responders ("Who")	1.73	0.78	<b>2.51</b>	1.36	0.45	<b>1.81</b>	0.17	0.12	<b>0.29</b>
Disaster outcomes	1.28	-	<b>1.28</b>	0.77	0.27	<b>1.04</b>	0.67	-	<b>0.67</b>
<b>Total</b>	<b>5.91</b>	<b>7.87</b>	<b>13.78</b>	<b>4.45</b>	<b>2.36</b>	<b>6.81</b>	<b>2.75</b>	<b>0.88</b>	<b>3.63</b>



Figure 3: Snapshot of video news coverage of the 2019 Notre Dame fire news cycle (an event in MultiVENT). The fire and the fallen spire were initially reported through first-person social media video uploads (at 10:02 AM and 10:51 AM, respectively) and then later broadcast by news organizations in more detail (10:38 AM, 11:10 AM, 12:20 PM, 12:26 PM). Some news coverage (12:20 PM) directly used first-person social media footage (10:51 AM). Hours later, news agencies uploaded more complete news stories with details and context (6:20 PM). This data suggests that it is important for models to learn from both first-person videos and official news coverage at various points in the news cycle to fully construct a factual model of the event, especially if the model is attempting to construct an event model while information develops online.

provided with the full set of videos  $V$  and a text query  $t \in T$ , and for each video  $v \in V$  returns the probability that  $t$  describes  $v$ , or  $h(v, t) = \mathbb{P}[f(v, t) = 1]$ . When there is a bijective mapping between queries and videos (e.g., when using video descriptions as queries), the model is evaluated on its recall when considering the top 1, 5, and 10 ranked videos (R@1, R@5, and R@10), as well as the median rank (MedR). When a given query may describe multiple videos, (e.g., when using event descriptions as queries), we instead evaluate the model on its precision given the top 1, 5, and 10 ranked videos (P@1, P@5, and P@10). We define these metrics as:

$$\text{Given } S := \arg \max_{V' \subseteq V: |V'|=k} \sum_{v \in V'} h(v, t),$$

Table 4: MultiCLIP evaluated alongside existing video retrieval approaches on the video retrieval benchmark MSR-VTT. Results indicate that MultiCLIP performs adequately on existing retrieval tasks, achieving comparable results to existing models. It does not perform as well as architectures that use multimodal transformers for joint encodings such as InternVideo and MPLUG-2.

Method	Year	Rank@1	Rank@5	Rank@10
FrozenInTime [4]	2021	32.5	61.5	71.2
Clip2Video [16]	2021	29.8	55.5	66.2
InternVideo [47]	2022	<b>55.2</b>	<b>79.6</b>	<b>87.5</b>
MPLUG-2 [50]	2023	53.1	77.6	84.7
<b>MultiCLIP</b>	2023	38.4	70.1	82.7

$$R@k = \frac{|\{s \in S : f(s, t) = 1\}|}{|\{v \in V : f(v, t) = 1\}|} \quad \text{and} \quad P@k = \frac{|\{s \in S : f(s, t) = 1\}|}{k}.$$

## 5.2 Model architecture and training

We introduce MultiCLIP, a multilingual baseline for video retrieval on MultiVENT. We base our architecture on the pretrained LAION CLIP ViT-H/14 frozen XLM-Roberta-Large model [11], which jointly trains an image and text encoder on text-image data to learn to pair images with their captions. At test time, it produces a zero-shot linear layer based on the test input’s visual features through which natural language captions can be passed in. The model architecture contains a vision encoder based on a ViT architecture [14] and a text encoder based on the the multilingual XLM Roberta large model [12]. A full overview of the CLIP architecture and pretraining can be found in the original paper [33].

In experiments using MultiCLIP, we first tokenize text descriptions using the XLM-Roberta-Large tokenizer, containing a vocabulary of over 250,000 words, and pass the tokens into MultiCLIP which produces a text embedding of size 1024. Next, we uniformly sample videos at a rate of 12 frames per video with an input size of 224x224, which the model uses to create a frame embedding of size 1024. To incorporate multilinguality into the model’s frame-level features, we use a ViT architecture trained with a contrastive objective over multilingual image-caption pairs from the LAION-5B dataset [42], which is constructed from the Common Crawl archive using images and their alt-text to produce a multilingual image-text dataset with over 100 languages. We mean pool the frame embeddings to produce a final video embedding, and use the text and video features to compute a similarity matrix of videos and descriptions.

## 5.3 Retrieval baselines

We first evaluate MultiCLIP on the existing video retrieval task MSR-VTT [51] using the recall metrics described in Sec. 5.1 alongside contemporary video retrieval approaches (FrozenInTime [4], Clip2Video [16], InternVideo [47], and MPLUG-2 [50]). Results on MSR-VTT’s validation set are reported in Table 4. The results indicate MultiCLIP performs well on standard video retrieval tasks, matching performance of separate text/vision pipeline models released within the last two years. It performs better than existing models that use separate text and vision pipelines (FrozenInTime [4] and Clip2Video [16]), but not as well as models that use larger architectures involving multimodal encodings (InternVideo [47] and MPLUG-2 [50]).

## 5.4 MultiVENT retrieval

We now evaluate MultiCLIP and related retrieval approaches on MultiVENT. We first use multilingual video descriptions as queries, and then we use English event summaries taken from the grounding text documents, meaning that one text query maps to up to ten videos. The event queries are selected by taking one to two sentences from each English event text document that describes the event most holistically. We exclusively use English queries for this section, as our annotators fluent in the other languages were not available for this task. In addition to MultiCLIP, we consider a set of contemporary video retrieval models with lightweight architectures (FrozenInTime [4], CLIP2Video



Table 5: Results showing the retrieval performance of video retrieval methods alongside MultiCLIP on MultiVENT. We use video descriptions and event descriptions as queries and partition results based on language. As shown, MultiVENT can be a difficult retrieval benchmark for video retrieval models even when considering only English, but the benefit of training on multilingual data is apparent when comparing MultiCLIP against the regular pooled CLIP model on non-English data.

Method	Video description				Event description		
	R@1	R@5	R@10	MedR	P@1	P@5	P@10
English							
FrozenInTime [4]	6.5	20.0	28.4	53.0	42.3	34.6	26.9
CLIP2Video [16]	41.3	71.8	80.4	2.0	96.2	<b>96.9</b>	73.3
InternVideo [47]	53.8	83.1	88.7	<b>1.0</b>	94.2	93.5	79.6
CLIP (pooled) [33]	<b>55.9</b>	83.9	91.3	<b>1.0</b>	98.1	94.6	<b>80.6</b>
<b>MultiCLIP</b>	<b>55.9</b>	<b>84.5</b>	<b>92.3</b>	<b>1.0</b>	<b>100.0</b>	<b>96.9</b>	<b>80.6</b>
Arabic + Chinese + Korean + Russian							
FrozenInTime [4]	0.5	1.2	2.5	793.5	29.8	22.6	17.6
CLIP2Video [16]	2.4	7.2	10.5	166.5	14.4	9.4	7.3
InternVideo [47]	5.7	13.9	19.8	91.0	79.3	71.0	55.7
CLIP (pooled) [33]	6.2	15.9	22.4	79.5	83.7	73.3	58.2
<b>MultiCLIP</b>	<b>32.6</b>	<b>64.7</b>	<b>79.5</b>	<b>3.0</b>	<b>85.6</b>	<b>76.4</b>	<b>61.5</b>

[16], InternVideo [47], and a pooled CLIP model using the same setup as MultiCLIP without the additional multilingual pretraining). We argue that lightweight architectures are most appropriate for evaluating a full, pairwise set of similarity scores between text and video data of large multimodal corpora. Results are reported, partitioned on language, in Table 5.

We report the standard recall @ rank  $k$  metric for retrieval on individual video queries, and precision @ rank  $k$  for retrieval on event description queries. The results suggest that some existing video retrieval models may particularly struggle on this task, regardless of language. We hypothesize that this is due to a combination of the videos’ length, complex semantic content, ambiguity, and frequent OCR content, as well as the long and often noisy video description queries.

While MultiVENT as a whole poses challenges to existing models, it is also clear that multilingual data may significantly impact performance on models trained primarily on English content - all models suffer a performance loss when evaluated on multilingual content (even when using English queries, as shown by the event description query results). While MultiCLIP suffers a performance loss on this data as well, comparing the standard pooled CLIP model against MultiCLIP shows that training on multilingual data does mitigate this multilingual performance loss: The two models perform comparably on English data, but MultiCLIP performs better on the multilingual content, especially when multilingual queries are used.

## 6 Conclusion

We introduce MultiVENT, a multimodal, multilingual dataset grounded in natural language documents for event-centric video retrieval and information acquisition. This dataset consists of 2,396 videos covering 260 current events reported in five target languages (Arabic, Chinese, English, Korean, and Russian) paired with multilingual natural language video descriptions and long-form event-centric text documents. We use this dataset to characterize online news coverage and how models can use this online content for information acquisition. We propose a multilingual video retrieval benchmark using MultiVENT and present MultiCLIP, multilingual video retrieval model to serve as a baseline for the task. We evaluate this model and related retrieval approaches on MSR-VTT and MultiVENT to illustrate the importance of pretraining on multilingual data for evaluation on MultiVENT. In future work, we aim to explore the effect that joint vision-OCR embeddings can have on video retrieval in text-heavy contexts. Also in future work, a RePAQ-adjacent system [25] for automatically extracting question-answer pairs from video content and video-document pairs could be developed and applied to MultiVENT. Through this, a framework for teaching models to perform open-domain question-answering tasks with multimodal background corpora could be established, expanding the domain of questions a model can answer.

## References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.
- [2] André Araujo, Jason Chaves, David Chen, Roland Angst, and Bernd Girod. Stanford i2v: a news video dataset for query-by-image experiments. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 237–242, 2015.
- [3] George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, et al. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. *arXiv preprint arXiv:2009.09984*, 2020.
- [4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.
- [5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker’s guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022.
- [6] Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006.
- [7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
- [8] Aozhu Chen, Ziyuan Wang, Chengbo Dong, Kaibin Tian, Ruixiang Zhao, Xun Liang, Zhanhui Kang, and Xirong Li. Chinaopen: A dataset for open-world multimodal learning. *arXiv preprint arXiv:2305.05880*, 2023.
- [9] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.
- [10] Wenhui Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.
- [11] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.
- [12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
- [13] Rishabh Dabral, Ganesh Ramakrishnan, Preethi Jyothi, et al. Rudder: A cross lingual video and text retrieval dataset. *arXiv preprint arXiv:2103.05457*, 2021.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [15] Joseph G Ellis, Brendan Jou, and Shih-Fu Chang. Why we watch the news: a dataset for exploring sentiment in broadcast video news. In *Proceedings of the 16th international conference on multimodal interaction*, pages 104–111, 2014.
- [16] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.
- [17] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022.
- [18] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group*, 8:1158–1178, 2010.
- [19] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. *arXiv preprint arXiv:2103.08849*, 2021.
- [20] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4441–4450, 2023.

- [21] Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326, 2011.
- [22] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.
- [23] Jie Lei, Tamara L Berg, and Mohit Bansal. mtvr: Multilingual moment retrieval in videos. *arXiv preprint arXiv:2108.00061*, 2021.
- [24] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.
- [25] Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenertorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115, 2021.
- [26] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, 2020.
- [27] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. *arXiv preprint arXiv:2005.02472*, 2020.
- [28] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.
- [29] Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. Narrative question answering with cutting-edge open-domain qa techniques: A comprehensive study. *Transactions of the Association for Computational Linguistics*, 9:1032–1046, 2021.
- [30] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.
- [31] Dina Pisarevskaya and Tatiana Shavrina. Wikiomnina: generative qa corpus on the whole russian wikipedia. *arXiv preprint arXiv:2204.08009*, 2022.
- [32] Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *arXiv preprint arXiv:2304.04358*, 2023.
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [34] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.
- [35] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- [36] Revant Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, et al. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11200–11208, 2022.
- [37] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. V3c—a research video collection. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*, pages 349–360. Springer, 2019.
- [38] Andrew Rouditchenko, Yung-Sung Chuang, Nina Shvetsova, Samuel Thomas, Rogerio Feris, Brian Kingsbury, Leonid Karlinsky, David Harwath, Hilde Kuehne, and James Glass. C2kd: Cross-lingual cross-modal knowledge distillation for multilingual text-video retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [39] Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute, 2016.

- [40] Piotr Rybak. Maupqa: Massive automatically-created polish question answering dataset. *arXiv preprint arXiv:2305.05486*, 2023.
- [41] Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. Ambiguous images with human judgments for robust visual event classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- [43] Christopher Sciaolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. *arXiv preprint arXiv:2109.08535*, 2021.
- [44] Jennifer Tracey, Ann Bies, Jeremy Getman, Kira Griffitt, and Stephanie Strassel. A study in contradiction: Data and annotation for aida focusing on informational conflict in russia-ukraine relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1831–1838, 2022.
- [45] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2022.
- [46] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.
- [47] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.
- [48] Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, et al. Newsnet: A novel dataset for hierarchical temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10669–10680, 2023.
- [49] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension. *arXiv preprint arXiv:2305.03347*, 2023.
- [50] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*, 2023.
- [51] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.
- [52] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641, 2003.
- [53] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.