This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MultiVENT 2.0: A Massive Multilingual Benchmark for Event-Centric Video Retrieval

Reno Kriz[♠♡*] Kate Sanders[♡*] David Etter[♠♡*] Kenton Murray[♠♡]
Cameron Carpenter[♡] Hannah Recknor[♠] Jimena Guallar-Blasco[♠]
Alexander Martin[♡] Eugene Yang[♠♡] Benjamin Van Durme[♠♡]
[♠] Human Language Technology Center of Excellence [♡] Johns Hopkins University
{rkriz1,ksande25,kenton}@jhu.edu

Abstract

*Efficiently retrieving and synthesizing information from large-scale multimodal collections has become a critical challenge. However, existing video retrieval datasets suffer from scope limitations, primarily focusing on matching descriptive but vague queries with small collections of professionally edited, English-centric videos. To address this gap, we introduce **MultiVENT 2.0**, a large-scale, multilingual event-centric video retrieval benchmark featuring a collection of more than 218,000 news videos and over 3,900 queries targeting specific world events. These queries specifically target information found in the visual content, audio, embedded text, and text metadata of the videos, requiring systems leverage all these sources to succeed at the task. Preliminary results show that state-of-the-art vision-language models struggle significantly with this task, and while alternative approaches show promise, they are still insufficient to adequately address this problem. These findings underscore the need for more robust multimodal retrieval systems, as effective video retrieval is a crucial step towards multimodal content understanding and generation.*

1. Introduction

While information retrieval systems for text documents have been extensively studied for decades, the landscape has shifted dramatically toward visual content, particularly videos. As of January 2024, YouTube alone likely hosts over 14 billion videos.[1] Despite this explosion of visual data, there remains a dearth of research focused on the efficient retrieval, processing, and synthesis of such vast collections. Datasets that reflect the diverse array of multimodal, multilingual news sources available online could help models adapt to this shift, but existing video retrieval datasets, including MSR-VTT [38], focus primarily on semantically simple videos created for English-speaking audiences. To address this limitation, the original MultiVENT dataset was introduced, containing **Multi**lingual **V**ideos of **E**vents with aligned **N**atural **T**ext across five languages [29]. However, both MULTIVENT (2,400 videos) and MSR-VTT (10,000 videos) are extremely small compared to standard information retrieval collections for text documents: In comparison, the HC4 corpus used in the 2022 NEUCLIR TREC shared task contains over 6 million text documents [17]. This disparity in collection size and diversity underscores the challenges in scaling video retrieval research.

To push the boundaries of video retrieval and create a more challenging and realistic task, we introduce MULTI-



Figure 1. Example query/video pairs from MSR-VTT [38] and MULTIVENT 2.0. MSR-VTT primarily contains broad descriptive queries mapped to general English-centric video clips, while MULTIVENT 2.0 targets specific current events covering 4 media formats, 6 languages, and subjects like natural disasters, politics, sports, social gatherings, and science.

[1] https://www.theatlantic.com/technology/archive/2024/01/how-many-videos-youtube-research/677250/

Dataset	Query Type	Query Generation	Domain	Source	# Queries	# Videos(/Clips)	Avg Clip Length	Multilingual
ActivityNetCaptions	Caption	Manual	Action	YouTube	100K	20K/100K	120s	no
MSR-VTT	Caption	Manual	Open	YouTube	200K	7.2K/10K	10-30s	no
VaTeX	Caption	Manual	Open/Act	Kinetics-600 (YouTube)	825K	42K/42K	10s	no
DiDeMo	Caption	Manual	Open	YFCC100M (Flickr)	40K	10.5K/26.9K	5s	no
LSMDC	Caption	Manual	Movie	Movies	128K	202/128K	4-5s	no
V3C	Metadata	-	Open	Vimeo	-	28K	8min	yes
MSVD	Caption	Manual	Open	YouTube	70K	2K/2K	4-10s	no
Valor-32K	Caption	Manual	Open	AudioSet (Youtube)	32K	32K/32K	10s	no
YouCook2	Caption	Manual	Cooking	YouTube	15.4K	2K/15.4K	20s	no
TVR	Query	Manual	TV Shows	TV Shows	109K	6/21.8K	76s	no
MTVR	Query	Manual	Movies	TV Shows	218K	6/21.8K	76s	yes
HowTo100M	Subtitle	Manual/Automatic	Instructional	YouTube	136M	1.2M/136M	4s	no
Multi-HowTo100M	Subtitle	Manual/Automatic	Instructional	YouTube	136M	1.2M/136M	9s	yes
MultiVENT 1.0	Caption	Manual	Open	YouTube/Twitter	2,400	2.4K	83s	yes
MULTIVENT 2.0 (ours)	Query	Manual	Open	YouTube/Twitter	3,900	218K	145s	yes

Table 1. Comparison of video retrieval datasets. Many of these datasets were collected as resources for general vision-language model training and did not explicitly target video retrieval. Subsequent use of these collections as a retrieval benchmark leveraged descriptive video captions or subtitles as proxy queries for English only data, which differ greatly in scope and difficulty from cross-lingual text retrieval datasets [17]. To our knowledge, MULTIVENT 2.0 is the first large-scale video dataset with queries explicitly developed for an event-focused multilingual retrieval task.

VENT 2.0. This dataset consists of more than 218,000 videos and over 3,900 manually-written queries targeting information about specific world events depicted within this video corpus. The videos primarily span six languages—Arabic, Chinese, English, Korean, Russian, and Spanish—and range from professionally edited news broadcasts to raw, first-person footage captured on cell phones. The events in MULTIVENT 2.0 include a wide array of types, including social and sporting events, disasters, political developments, and scientific discoveries. Queries were designed to target specific information from the visual content, audio, embedded text, and text metadata, which challenges models to effectively process and integrate information across modalities and languages and better reflects real-world retrieval scenarios.

Preliminary results show that this task presents significant challenges for current state-of-the-art vision-language models (VLMs). While specialized single-modality models show promise for queries targeting content from their respective modalities, they remain insufficient for addressing the full range of queries. These findings suggest that existing systems are not yet equipped to handle complex vision-language tasks and underscores the need for more robust multimodal systems, as effective video retrieval is a critical step toward multimodal content understanding and generation. The main contributions of this paper are:

1. We introduce MULTIVENT 2.0,[2] a large-scale multilingual video retrieval task containing more than 218,000 videos. Over 3,900 queries were crafted to target aspects of current events using visual content, audio, embedded text, and text metadata from these videos.
2. We evaluate a variety of video retrieval models, demonstrating that even state-of-the-art multimodal systems struggle in this challenging event-centric setting.
3. We baseline specialized modality-specific models; despite not being able to address the task on their own, we show their potential usefulness as components in more robust retrieval pipelines.

[2]Data and intermediate features are available on Huggingface.

2. Related Work

Video Retrieval Datasets There are an increasing number of video datasets, many of which support the downstream text-video retrieval task. However, there are various limitations to these existing datasets. Many of these dataset have too few videos or have videos that are too short for comprehensive evaluation of complex event understanding [28]. For example, the popular baseline dataset, MSVD [3] only has 1,970 videos whose length average between 4 and 10 seconds. The larger and more recent VATEX [32] and VALOR-32K datasets [5] have video clips that are on average only 10 seconds long, and the clips of the extremely large dataset of over 1 million videos, HOWTO100M [23], average to only 4 seconds. Many video dataset are limited to videos of a single topic or domain, such as the LSMDC dataset [26] and TVR dataset [18], which feature clips from movies and TV shows respectively, and the YOUCOOK2 dataset which features cooking videos [42]. Thus, while these datasets are quite large in terms of number of videos, they do not represent diverse events or "real-world" topics. Moreover, while some datasets do have non-English videos, they are limited to only one non-English language; MULTI-HOWTO100M [14] and MTVR [19] both have only English and Chinese videos. Furthermore, despite the increase in Video datasets, very few video datasets can support event-retrieval or true multi-modal retrieval. The DIDEMO

dataset [12] is a large dataset featuring event-centric videos. However, the creators of the DiDeMo dataset filtered out any non-professional edited videos, thus limiting the scope of the dataset. Additionally, captions or video descriptions are often extracted for pretraining as the "query" intended for the text-video retrieval task such as ActivityNet captions [16] and MSR-VTT [38]. The V3C dataset [27] does not use captions or subtitles but video metadata. While video descriptions or summaries are high-level representations of a video, they do no reflect human style search queries befitting of a realistic retrieval task. To more easily compare and contrast the existing datasets, we have compiled the details of said video datasets in Table 1.

Video Retrieval Methods Text-video retrieval is a core research area in video-language understanding [1, 4, 7, 31, 33, 34, 40, 41]. There have been many proposed solutions focusing on: combining pre-extracted features from frozen text and vision encoders [41]; transferring contrastive pretraining architectures, such as CLIP [24], to the video domain [10, 21, 22, 39]; leveraging video specific features such as sparsity [1, 20]; and aligning/fusing modalities [5, 6, 13, 15, 35, 37, 43]. Emphasis has been placed on aligning visual content with text, yet videos also include valuable information from text metadata, audio, and embedded text. Recently, fusion models [5, 6] and binding models [11, 43] have been developed to unify representations across modalities. VAST [6] notably integrates vision, audio, and subtitle information for text-video retrieval, enabling higher semantic video understanding in a unified representation. Despite these advances, many of these models are trained on datasets not designed for single-modality retrieval, let alone retrieval across multiple modalities. As a result, these systems face challenges when handling more complex retrieval tasks requiring inference across modalities and languages. This highlights the need for more comprehensive collections to facilitate the development of more robust and adaptable video retrieval systems capable of addressing the diversity of events found in real-world multimedia content.

3. Video Collection

Two significant challenges in current video retrieval research are: (1) existing collections often reward models for scene description alone, without considering the broader event context, and (2) these collections are typically limited in both size and scope. In this section, we outline the video collection process for MultiVENT 2.0, demonstrating how it addresses these challenges. We first review the development of MultiVENT 1.0, which initiated efforts to tackle the first issue, and then describe the expanded collection process for MultiVENT 2.0, a larger and more diverse dataset designed for real-world applications. The final MultiVENT 2.0 dataset comprises over 217,000 videos, split evenly into train and test collections.

3.1. MultiVENT 1.0 Development and Limitations

The creation of MultiVENT 1.0 involved several key steps: developing topics based on language- and country-specific current events, collecting relevant videos for each topic, and aligning these events with corresponding news articles. For topic selection, Sanders et al. [29] utilized Google Trends statistics from countries with the largest populations of speakers for each target language to identify visually salient current events. After filtering to ensure adequate online video coverage, this process produced 2,396 videos spanning 255 events.

This principled approach resulted in an important dataset, filling a unique role as a targeted, event-centric video retrieval collection. However, while MultiVENT 1.0 was a significant step forward, its small size, especially the limited distractor set, limits its applicability for large-scale multimodal retrieval research. As shown in Table 1, this issue is common across many video retrieval tasks, and stands in stark contrast to text retrieval benchmarks, which typically consist of much larger collections. For example, the HC4 dataset used in the 2022 NeuCLIR TREC shared task contains over 6 million text documents [17]. This disparity in collection size underscores the need for a larger and more diverse video retrieval corpus.

3.2. Expanded MultiVENT 2.0 Video Collection Process

To create an expanded collection for a more realistic, large-scale video retrieval task, we augment the original MultiVENT dataset with videos from InternVid, a corpus containing more than seven million YouTube videos and over 760,000 hours of content [36]. InternVid covers a superset of target languages and event categories than those found in MultiVENT 1.0. While not all InternVid categories are event-based, the corpus still includes a significant amount of event-centric content, particularly in the political and disaster domains.

For MultiVENT 2.0, we extract a large subset of videos from InternVid, filtering out those longer than five minutes. This process yields approximately 40,000 videos for each of the MultiVENT 1.0 target languages: Arabic, Chinese, English, Korean, and Russian. To introduce greater variability at test time, we also include a smaller set of videos from the Spanish and "Unknown" language categories, ensuring that systems handle both new primary languages and a long tail of low-resource languages.

After merging the expanded collection with MultiVENT 1.0, we compiled a final dataset of over 217,000 videos, consisting of 108,500 videos for training (MultiVENT Train) and 109,800 for testing

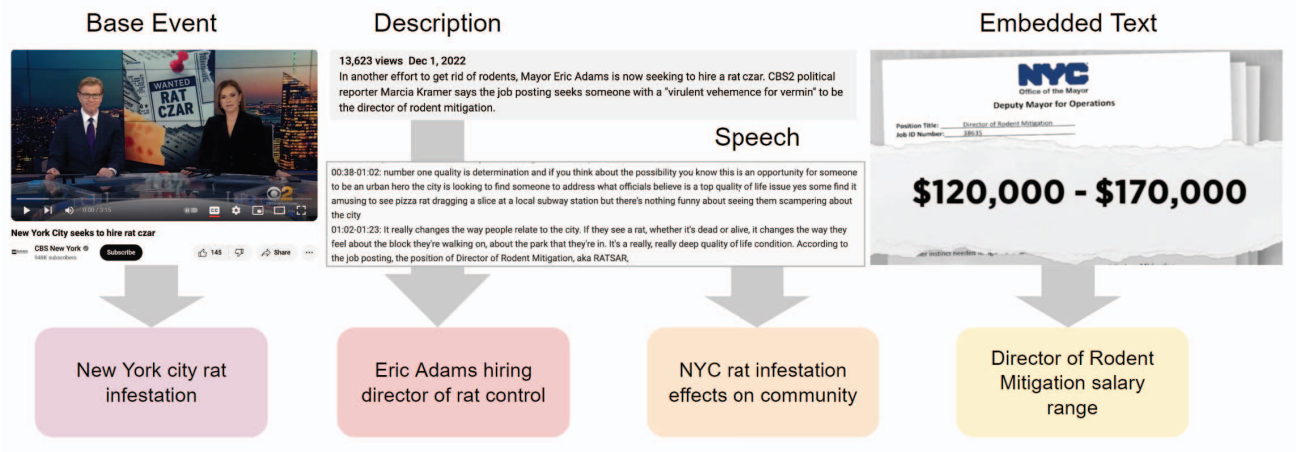


Figure 2. Query creation process for event-centric videos within our distractor collection from InternVid. Annotators first create a **Base Event** query based on the primary event depicted in the video. They then write up to three additional queries focusing on specific and unique aspects of the event: the **Description** query uses only information from the human-written text description, the **Speech** query relies on spoken content from the video, and the **Embedded Text** query utilizes text visible within the video frames.

(MULTIVENT TEST). All videos from MULTIVENT 1.0 are solely found in the evaluation set, and any duplicates between the training and test collections were removed. Given the scale of the dataset, rapidly comparing models and system variations can be challenging. To mitigate this and aid model tuning, we offer a subset of 2,000 videos from the training set, referred to as MULTIVENT TRAIN-2K.

4. Query Creation

As mentioned in Section 3, prior collections have focused more on matching descriptive aspects within videos. Accordingly, search queries have either been relatively short and vague (e.g., *a black and white horse runs around*) or derived directly from a video's metadata, such as a YouTube description. MULTIVENT 1.0 largely followed this precedent, relying on the available metadata for each video. However, these approaches create a disconnect with modern text retrieval practices, where concise, search-engine-style English queries are increasingly used to match multilingual documents [17]. To bridge this gap, we develop a novel two-pronged approach for creating event-centric video queries: one that leverages additional fine-grained event annotations available with MULTIVENT 1.0 videos [29], and another that relies solely on the videos themselves.

For this task, we recruit professional linguists with expertise across the six primary languages targeted in MULTIVENT 2.0 to develop the queries. Each annotator underwent training via a tutorial task with detailed instructions. Afterward, we provide one-on-one feedback on their performance and closely monitor their initial annotations, offering additional guidance as needed. Finally, an annotation lead conducts a quality control check to ensure the conciseness and specificity of each query/video pair.

4.1. Updated Query Creation for MultiVENT 1.0

In MULTIVENT 1.0, the text descriptions accompanying each video were initially repurposed as queries. However, this dataset also provides a foundation for crafting more targeted, event-centric queries, as each of the 255 current events in the collection is aligned with an article in its original language, along with an English version if the source language is not English. Many of these English articles come from Wikipedia, where titles are typically concise and specific descriptions of the broader event. For instance, the article aligned with the event shown in Figure 1 is titled *2022 Lotus Garden China Telecom Building fire*. Annotators follow this style to manually create a base query for each current event in MULTIVENT 1.0, henceforth referred to as MULTIVENT BASE event queries.

While retrieving videos about overarching events is a valuable step forward, research in text information extraction (IE) has shown that individual events often encompass multiple distinct aspects, which are unlikely to be fully captured in a single piece of media. Thus, to further challenge retrieval models, annotators develop queries focused on specific, unique aspects of each event. This is facilitated by leveraging annotations from MULTIVENT-GROUNDED [30], a collection of 1,200 MULTIVENT 1.0 videos containing fine-grained event annotations. These annotations, guided by adapted FRAMENET event templates, include text description spans, video time intervals, and spatial bounding boxes at the frame level, and address IE-based questions such as *Where did the disaster occur?*, *Who was affected by the disaster?*, and *Who responded to the disaster?* Based on these annotations, annotators create 884

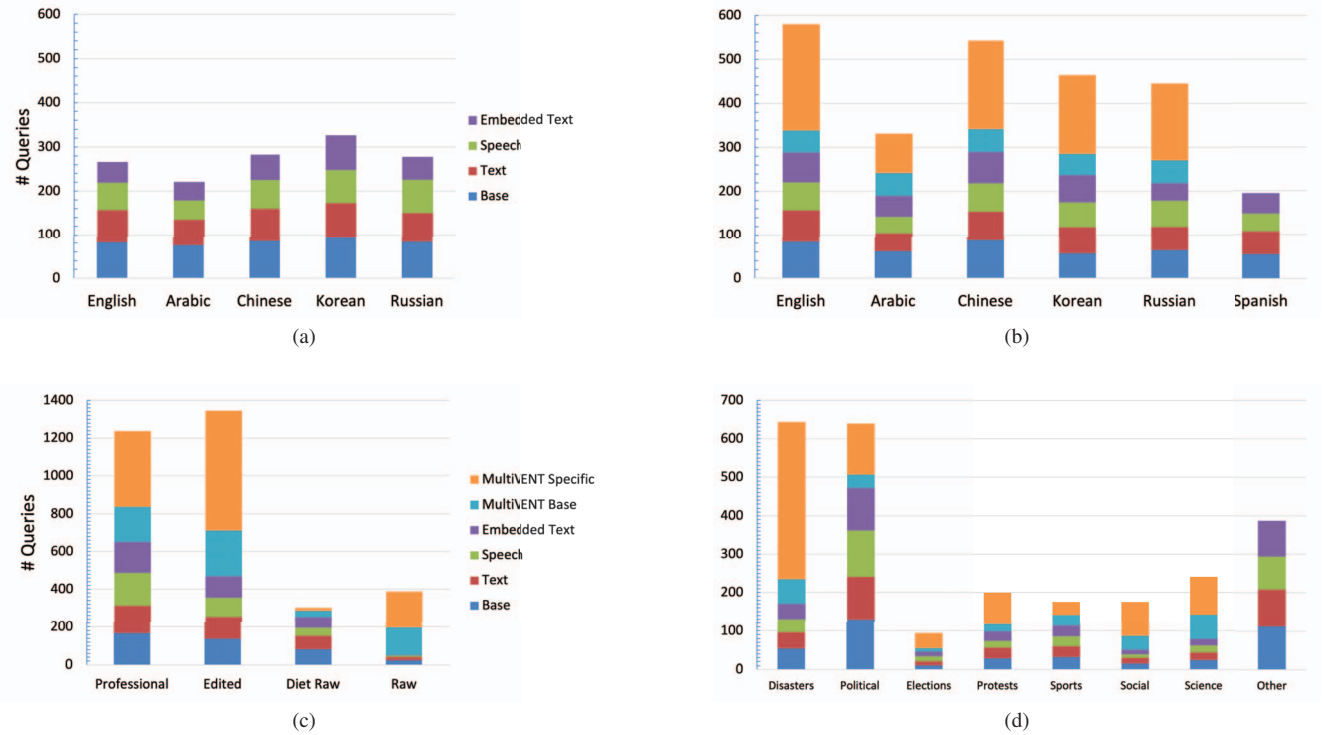


Figure 3. Breakdowns of of the number of queries mapped to relevant videos. Figure 3a shows that MULTIVENT TRAIN contains queries targeting the five primary languages from MULTIVENT 1.0. On the other hand, in Figure 3b we see that MULTIVENT TEST adds queries targeting Spanish events to challenge systems' multilingual robustness. Figure 3c shows that MULTIVENT 2.0 targets videos ranging from professional news broadcasts to raw first-person footage of events. Finally, as seen in Figure 3d, events in MULTIVENT 2.0 generally map to the same categories as MultiVENT 1.0, with a long tail of infrequent event types.

MULTIVENT SPECIFIC queries, each targeting videos that highlight distinct aspects of the corresponding event.

4.2. InternVid Query Creation

The previous section focuses on the retrieval of MULTIVENT 1.0 videos, despite the majority of our collection coming from InternVid. This raises the risk that part of the task could devolve into a binary classification problem, where simply distinguishing between MultiVENT and InternVid videos significantly reduces the task's complexity. To mitigate this, we tasked annotators with writing queries targeting new events found within the larger InternVid collection. Since there was no guarantee that InternVid videos contained event-based content, we focused annotation efforts on videos from the News & Politics and Sports categories. Annotators were first asked to confirm whether a video contained event-based content. If confirmed, they write a **Base Event** query in the same style as those described in Section 4.1. Additionally, annotators searched the web for an article related to the event.

Next, annotators develop up to three additional queries, each focusing on specific and unique aspects of the same event. For these queries, annotators were asked to rely on partial information about a video from a single modality: the **Description** query uses only the human-written YouTube description of the video; the **Speech** query uses any spoken content from the video, with Whisper-generated speech transcripts provided for assistance [25]; and the **Embedded Text** query focused solely on text directly visible in the video frames, with output provided from a multilingual OCR system [9]. Note that not all videos contain useful or unique information across all modalities; in such cases, no query is created. This process was applied to both the train and test collections, resulting in 1,417 additional test queries and 1,375 train queries.

4.3. Query and Video Breakdowns

To gain insight into the challenges posed by different types of video/query pairs, annotators are asked several additional questions about the videos during the query creation process. First, we ask them to confirm the video's primary language. While INTERNVID ostensibly provides this information, the labels are sometimes incorrect; this is particularly common in English-labeled videos, where approximately 20% of the videos were in other languages. Figure 3a shows the distribution of queries targeting videos

from each of the five original MULTIVENT 1.0 languages within MULTIVENT TRAIN, while Figure 3b illustrates the same for MULTIVENT TEST. Notably, we add Spanish as an additional primary language in the test collection to evaluate the multilingual robustness of video retrieval systems.

Additionally, linguists group events into one of the seven event types defined in MULTIVENT 1.0: natural disasters (**Disasters**), political elections (**Elections**), **Protests**, other political developments (**Political**), sporting events (**Sports**), social events (**Social Events**), and scientific or technological discoveries (**Science**) [29]. An **Other** category is also provided for events that did not fit into these types. Figure 3d depicts the event type breakdown. Two notable aspects include the relative abundance of MultiVENT-specific queries for disaster events, largely due to MULTIVENT-GROUNDED annotations prioritizing this category [30], and the predominance of Political events, which reflects the news and politics videos comprising the majority of event-based content within the InternVid collection. The Other category includes diseases, police incidents, and man-made disasters, among other infrequent event types.

Finally, we ask annotators to categorize videos into three general types: **Professional** news broadcasts, i.e., videos with reporters and/or traditional news chyron; non-professional **Edited** videos, i.e., videos featuring multiple spliced clips, visual effects, or superimposed graphics; and **Raw** footage, i.e., single-stream videos of events as they happen, typically captured on a mobile device. Given the widespread availability of video editing software, it is very easy for more people to make light edits prior to uploading. To differentiate between this and "true" raw content, we add a fourth category, **Diet Raw**, for single-stream videos with minimal text and speech overlays. Figure 3c shows the distribution of queries for relevant videos across these types. From this, we can see that true raw content is relatively scarce in InternVid, reinforcing the importance of the original MultiVENT collection. Note that many queries targeting MultiVENT 1.0 events are mapped to multiple relevant videos across different video types.

5. Relevance Judgment Annotation

After the initial query creation process, we have a total of 6,068 annotated video/query pairs for MULTIVENT-TEST. In many cases, queries are only mapped to a single relevant video, and all other non-judged videos are assumed to be not relevant. This presents a challenge because, despite the targeted nature of our queries, it is likely that the distractor set contains additional relevant videos, given the size and diversity of the test collection. Furthermore, some videos may be partially relevant to a query, or relevant to some parts but not all, while others' relevance may be unclear due to a lack of context or inherent video ambiguity.

To address this, we re-train the same pool of professional linguists from the query creation task in Section 4 to now judge the relevance of previously unseen query/video pairs. For each candidate video, annotators are asked to classify it as *not relevant*, *possibly relevant*, *partially relevant*, or *very relevant* to a query. Given the potentially vast scope of this annotation task, we limit the videos to those ranked in the top 10 by multilingual CLIP (MCLIP), a pre-trained vision-language model [2] and the strongest-performing model on MULTIVENT 1.0 [29]. After removing video/query pairs already judged, we prioritize judging the highest-ranked videos, and streamline the process by grouping query/video pairs where the same video was ranked in the top 10 for multiple queries. Through this process, we collect an additional 4,396 gold relevance judgments. For evaluation purposes, we condense the middle two categories into a single category labeled *somewhat relevant*.

Rank	Very Relevant	Somewhat Relevant	Not Relevant
1	43%	13%	44%
2	28%	15%	57%
3	28%	19%	53%
4	20%	21%	59%
5	18%	18%	65%
6	14%	18%	69%
7	14%	17%	70%
8	10%	11%	78%
9	10%	15%	75%
10	9%	13%	79%

Table 2. Percentage of initially un-judged video/query pairs re-annotated as not relevant, somewhat relevant, and very relevant. The *Rank* indicates the ranking of a video for the corresponding query, as judged by our best baseline model.

Given these graded relevance categories, it follows that any video judged as very relevant for a **MultiVENT Base** query must be at least somewhat relevant for all associated **MultiVENT Specific** queries. To ensure systems receive appropriate credit, we apply this logic to add an additional 5,653 silver relevance judgments. This results in a final updated set of 16,116 judged video/query pairs. Prior to this annotation effort, on average only 16% of the videos ranked in the top 10 by multilingual CLIP had an associated relevance judgment (*Judged@10*); with the added judgments, the *Judged@10* has now increased to 39%. For evaluation metrics that account for graded relevance, such as normalized Discounted Cumulative Gain (nDCG), we apply a 0-1-3 scale: *Very relevant* video/query pairs are scored as 3, *Somewhat Relevant* pairs as 1, and *Not Relevant* or non-judged pairs as 0.

6. Baselines and Results

Jointly pre-trained vision-language models (VLMs) have recently achieved state-of-the-art results on various downstream video understanding tasks. These models typically

Modality	Model	MSR-VTT	MultiVENT 2.0				
		R@10	R@10	R@100	MRR	mAP	nDCG@10
Vision	mCLIP	0.827	0.333	0.603	0.429	0.261	0.303
OCR	ICDAR → mCLIP	-	0.227	0.374	0.363	0.166	0.217
Speech	Whisper → mCLIP	-	0.290	0.450	0.417	0.212	0.267
Text	Description → mCLIP	-	0.293	0.491	0.445	0.228	0.284
Vision	InternVideo2.0	**0.851**	0.004	0.018	0.018	0.003	0.005
All	VAST	0.739	0.118	0.118	0.198	0.080	0.116
All	LanguageBind	0.787	**0.355**	**0.620**	**0.443**	**0.283**	**0.324**

Table 3. Performance of pre-trained multimodal benchmark models and single-modality pipeline systems. We can see that prior state-of-the-art video retrieval systems struggle significantly with the increase in query complexity and collection size/diversity of MULTIVENT 2.0. VAST is run with vision input only on MULTIVENT 2.0, as that achieves highest performance on this task.

consist of separate single-modality encoders, with a mechanism for fusing multimodal embeddings during training. In this section, we consider several prominent VLMs as baselines for video retrieval. Each of the following VLMs utilize a version of CLIP for the vision component, with the notable exception of INTERNVID using a Vision Transformer (ViT) [8].

- **VALOR** combines three encoders for single-modality representations with a decoder designed for multimodal text generation [5].
- **VAST** utilizes omni-modality pretraining to simultaneously fine-tune text, audio, and image encoders to enhance cross-modality learning [6].
- **InternVid 2** employs two expert models for video token-level unmasking during training and currently is state-of-the-art for zero-shot retrieval on MSR-VTT [37].[3]
- **Language-Bind** leverages a CLIP-based encoder for all non-text modalities and has achieved state-of-the-art results on several video-understanding tasks [43].

Beyond VLMs, we also evaluate several baseline pipeline approaches that utilize information from a single modality. In all cases, these system leverages MCLIP's multimodal embedding space to efficiently compare model outputs to queries.

- **mCLIP**: We first extract ten keyframes by detecting significant scene changes within a video and then extracting the midpoint frame.[4] The frames are passed through MCLIP's image encoder, and the pooled embeddings are then compared to analogous query embeddings.
- **ICDAR** OCR: Using the same frames from the vision baseline, we extract all visible embedded text using a state-of-the-art multilingual OCR system [9], and pass the text output through mCLIP's text encoder.
- **Whisper** OCR: We utilize WHISPER, a robust automatic speech recognition system [25], to transcribe each video's audio track, and the resulting text is passed through MCLIP's text encoder.
- **Description**: This baseline embeds each video's human-written description through MCLIP's text encoder.

Table 3 presents the performance of these baselines on MULTIVENT 2.0, as well as results on MSR-VTT, a standard benchmark video retrieval datasets. A key observation is that while VLMs excel on prior collections, most perform poorly on our task. This is likely due to two main factors: the length of our videos, as VLMs are typically trained on short video segments; and the significant domain mismatch between simple visual concepts and complex event-based natural language, the latter being much harder to map directly to visual features. This second issue is compounded by some MULTIVENT 2.0 queries targeting non-visual aspects of the video.

Another notable takeaway is that single-modality pipeline systems show promise, though there is considerable room for improvement. Table 4 highlights how different aspects of the problem can be better addressed by specific modalities. The distinction is particularly evident in the breakdown of video types in Table 4b: speech is highly effective for retrieving professional news broadcasts, yet it, along with embedded text, prove to be almost useless with raw content. In contrast, human-written descriptions and visual content are more robust across video types, with descriptions being the most useful for retrieving raw footage.

Based on these findings, a key question moving forward is whether to allow the use of text metadata. This may be appropriate in certain settings, but given the rapid rise of online visual content, it is unrealistic to assume that clean, human-written descriptions of a video will always be available. This is especially true for raw, user-generated content, which is often posted in real-time as events unfold.

[3] https://paperswithcode.com/sota/zero-shot-video-retrieval-on-msr-vtt

[4] https://www.scenedetect.com

Modality	Model	Performance by Language					
		Arabic	Chinese	English	Korean	Russian	Spanish
Vision	mCLIP	**0.323**	0.074	**0.331**	0.070	**0.247**	0.272
OCR	ICDAR → mCLIP	0.320	0.083	0.171	0.107	0.104	0.289
Speech	Whisper → mCLIP	0.232	**0.130**	0.203	0.114	0.224	**0.320**
Text	Description → mCLIP	0.246	0.127	0.199	**0.137**	0.242	0.281

(a)

Modality	Model	Performance by Video Type			
		Professional	Edited	Diet Raw	True Raw
Vision	mCLIP	0.201	**0.276**	**0.152**	0.198
OCR	ICDAR → mCLIP	0.184	0.176	0.094	0.006
Speech	Whisper → mCLIP	**0.318**	0.154	0.011	0.021
Text	Description → mCLIP	0.206	0.245	0.142	**0.307**

(b)

Modality	Model	Performance by Query Type					
		MultiVENT-Base	Multivent-Specific	Base	Text	Speech	Embedded Text
Vision	mCLIP	**0.411**	0.331	**0.294**	**0.294**	0.227	0.232
OCR	ICDAR → mCLIP	0.254	0.198	0.219	0.232	0.162	**0.245**
Speech	Whisper → mCLIP	0.268	0.225	0.282	0.256	**0.306**	0.230
Text	Description → mCLIP	0.403	**0.359**	0.219	0.282	0.160	0.153

(c)

Modality	Model	Performance by Event Type							
		Disaster	Election	Protest	Political	Sports	Social	Science	Other
Vision	mCLIP	0.334	**0.215**	**0.310**	**0.334**	**0.441**	**0.486**	**0.373**	0.259
OCR	ICDAR → mCLIP	0.208	0.153	0.244	0.186	0.284	0.287	0.228	0.218
Speech	Whisper → mCLIP	0.242	0.111	0.221	0.290	0.267	0.214	0.320	**0.321**
Text	Description → mCLIP	**0.343**	0.173	0.260	0.227	0.335	0.443	0.302	0.211

(d)

Table 4. **Breakdown of single-modality pipeline results.** Table 4a shows that Chinese and Korean videos are particularly challenging, especially for vision-only models. From Table 4b, extracted text is most beneficial for professional news broadcasts, while real-time raw footage remains challenging for all direct video content. Table 4c demonstrates that queries based on specific modalities are generally best handled by the corresponding modality. Finally, from Table 4d, visual content is most effective for most MultiVENT 1.0 event types.

As a result, we propose two versions of the final evaluation task for future systems: one where descriptions are disallowed (MULTIVENT TEST-NODESC) and one where they are permitted (MULTIVENT TEST-DESC).

7. Conclusion

In this paper, we introduce MULTIVENT 2.0, a massive multilingual, event-centric video retrieval collection that significantly broadens the scope and complexity of multimodal retrieval tasks. With over 218,000 videos and 3,900 queries targeting a diverse range of world events across six languages, MULTIVENT 2.0 poses substantial new challenges for vision-language models. Our results indicate that while pre-trained models such as InternVid and VAST have achieved impressive performance on prior retrieval collections, they face considerable difficulties in this setting, largely due to the domain mismatch between interpreting simpler visual concepts and the complex multimodal nature of event-based content. Pipeline systems incorporating specialized single-modality models show promise for subsets of the task, but these remain insufficient in isolation. In addition, retrieving raw video footage without the accompanied human-written text descriptions is extremely challenging, a limitation especially relevant in real-time event scenarios. Our results highlight the need for more robust multimodal retrieval systems, as exciting visual understanding and generation tasks cannot be leveraged in practice without being able to accurately and efficiently triage relevant visual content.

8. Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE2139757. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. Rap: Efficient text-video retrieval with sparse-and-correlated adapter, 2024. 3

[2] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, 2022. European Language Resources Association. 6

[3] David Chen and William 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, Portland, Oregon, USA, 2011. Association for Computational Linguistics. 2

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning, 2020. 3

[5] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset, 2023. 2, 3, 7

[6] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset, 2023. 3, 7

[7] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval, 2021. 3

[8] 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, 2021. 7

[9] David Etter, Cameron Carpenter, and Nolan King. A hybrid model for multilingual ocr. In *Document Analysis and Recognition - ICDAR 2023*, pages 467–483, Cham, 2023. Springer Nature Switzerland. 5, 7

[10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip, 2021. 3

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3

[12] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language, 2017. 3

[13] Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a unified video-language alignment and fusion model, 2022. 3

[14] 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, 2021. 2

[15] Peng Jin, JinFa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *Advances in Neural Information Processing Systems*, pages 30291–30306, 2022. 3

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos, 2017. 3

[17] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. Overview of the trec 2022 neuclir track, 2023. 1, 2, 3, 4

[18] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval, 2020. 2

[19] Jie Lei, Tamara L. Berg, and Mohit Bansal. Mtvr: Multilingual moment retrieval in videos, 2021. 2

[20] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling, 2021. 3

[21] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval, 2021. 3

[22] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 638–647, New York, NY, USA, 2022. Association for Computing Machinery. 3

[23] 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. 2

[24] 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, 2021. 3

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 5, 7

[26] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description, 2016. 2

[27] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. V3c - a research video collection, 2018. 3

[28] Kate Sanders and Benjamin Van Durme. A survey of video datasets for grounded event understanding. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7314–7327, 2024. 2

[29] Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. Multivent: Multilingual videos of events with aligned natural text, 2023. 1, 3, 4, 6

[30] Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. Grounding partially-defined events in multimodal data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15905–15927, Miami, Florida, USA, 2024. Association for Computational Linguistics. 4, 6

[31] Haoran Tang, Meng Cao, Jinfa Huang, Ruyang Liu, Peng Jin, Ge Li, and Xiaodan Liang. Muse: Mamba is efficient multi-scale learner for text-video retrieval, 2024. 3

[32] 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, 2020. 2

[33] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval, 2021. 3

[34] Yimu Wang and Peng Shi. Video-text retrieval by supervised sparse multi-grained learning, 2023. 3

[35] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022. 3

[36] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024. 3

[37] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024. 3, 7

[38] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 1, 3

[39] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment, 2023. 3

[40] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment, 2021. 3

[41] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval, 2018. 3

[42] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[43] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2024. 3, 7