6-2024

Improving interpretable embeddings for ad-hoc video search with generative captions and multi-word concept bank

Jiaxin WU

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Wing-Kwong CHAN

Citation

WU, Jiaxin; NGO, Chong-wah; and CHAN, Wing-Kwong. Improving interpretable embeddings for ad-hoc video search with generative captions and multi-word concept bank. (2024). *ICMR '24: Proceedings of the 2024 International Conference on Multimedia Retrieval, Phuket, Thailand, June 10-14*. 73-82.
Available at: https://ink.library.smu.edu.sg/sis_research/9288

Improving Interpretable Embeddings for Ad-hoc Video Search with Generative Captions and Multi-word Concept Bank

Jiaxin Wu
jiaxin.wu@my.cityu.edu.hk
City University of Hong Kong
Hong Kong

Chong-Wah Ngo
cwngo@smu.edu.sg
Singapore Management University
Singapore

Wing-Kwong Chan
wkchan@cityu.edu.hk
City University of Hong Kong
Hong Kong

ABSTRACT

Aligning a user query and video clips in cross-modal latent space and that with semantic concepts are two mainstream approaches for ad-hoc video search (AVS). However, the effectiveness of existing approaches is bottlenecked by the small sizes of available video-text datasets and the low quality of concept banks, which results in the failures of unseen queries and the out-of-vocabulary problem. This paper addresses these two problems by constructing a new dataset and developing a multi-word concept bank. Specifically, capitalizing on a generative model, we construct a new dataset consisting of 7 million generated text and video pairs for pre-training. To tackle the out-of-vocabulary problem, we develop a multi-word concept bank based on syntax analysis to enhance the capability of a state-of-the-art interpretable AVS method in modeling relationships between query words. We also study the impact of current advanced features on the method. Experimental results show that the integration of the above-proposed elements doubles the R@1 performance of the AVS method on the MSRVTT dataset and improves the xinfAP on the TRECVid AVS query sets for 2016–2023 (eight years) by a margin from 2% to 77%, with an average about 20%.

KEYWORDS

Ad-hoc video search, Interpretable embedding, Large-scale video-text dataset, Concept bank construction, Out of vocabulary

1 INTRODUCTION

With the ever-growth of video data, e.g., videos sharing on YouTube or TikTok, text-to-video search is an essential tool for users to find videos of their interests. Especially, ad-hoc video search (AVS), which allows users to retrieve videos through an open-vocabulary textual query, has attracted lots of attention in many years [23, 31, 48, 58, 63]. AVS is challenging as it does not provide any annotated data for training, and the test collection is usually huge (e.g., V3C1 dataset [8], which has over 1 million video clips).

Concept-based search [9, 37, 58] and embedding-based search [11, 14, 30, 31] are two mainstream approaches for ad-hoc video search. Concept-based search indexes a user query and video clips by semantic concepts. Specifically, the concepts of the video clips are extracted by classifiers trained on image/video classification datasets. The query text is mapped to concept tokens of a concept bank through a concept selection. The effectiveness of a concept-based search relies on the qualities of the concept selection and the concept bank, as well as the accuracies of the video concept extractors. In contrast, by getting rid of the tedious concept extraction and selection, embedding-based search (aka concept-free search) aligns a user query with video clips in a joint latent space. With more video caption datasets available, embedding-based search has outperformed concept-based search and has become a mainstream approach for AVS since 2018. However, since the alignment is conducted on the high-dimensional latent space, the result of the embedding-based search is not explainable and predictable. To this end, the dual-task model [62] proposes to interpret the joint latent space with semantic concepts and align a user query with video clips in an interpretable embedding space. Some recent studies [16, 63] follow the effort of the dual-task model to build interpretable approaches. For example, the ITV model [63] proposes more consistent interpretations for query-video embeddings with unlikelihood training, and the RIVRL model [16] aligns a user query and videos in both a latent space and a concept space. As concept-based search and embedding-based search are complementary, the interpretable approaches are promoted as a new state-of-the-art for the AVS task.

Nevertheless, the existing AVS approaches are bottlenecked by small video-text datasets for training and usually fail on unseen queries. On the one hand, some approaches try to take advantage of large-scale image-text datasets and utilize text-to-image retrieval models for text-to-video retrieval [59, 60]. To some extent, the text-to-image retrieval models work for small-size datasets, e.g., MSRVTT dataset [65] with 10k videos. However, when the dataset size becomes huge, such as over 1 million video clips, either text-to-image retrieve models or their extension to video level, e.g., CLIP [45] or CLIP4CLIP [36], is not competitive to a state-of-the-art AVS model trained on small-size video caption datasets as evidenced in the TRECVid AVS evaluations [2–4]. On the other hand, some huge text-video datasets (e.g., HowTo100M [38], and WebVid [7]) are released to facilitate the learning of text and video alignments. However, existing large-scale datasets only have one caption per video, which overlooks rich semantics in videos. As a query text is usually ad-hoc, having multiple versions of video captions is essential for learning robust query-video alignments. In addition to the small-size dataset problem, the existing approaches only leverage the word-only concept banks for interpretation [16, 63]. The word-only concept bank is inherently limited in modeling the word-to-word relationship, resulting in out-of-vocabulary in modeling word composition, such as those with prepositional words (e.g., "in front of").

In this paper, we follow the effort of the interpretable approaches and propose three general and feasible components to address the dataset size and out-of-vocabulary problems: (1) Capitalizing on a generative model, we construct a new large-scale text-video dataset automatically for video-text retrieval models without human effort. The dataset (named WebVid-genCap7M) has 7 million generated text-video pairs, which is available online[1], and it is used for the

[1] WebVid-genCap7M dataset

retrieval model pre-training. (2) We develop a multi-word concept bank to address the out-of-vocabulary problem in concept-based search by incorporating various phrases based on syntactic analysis of sentences. The evaluation shows that the new concept bank is able to bring an average of about 60% gain of xinfAP to a state-of-the-art concept-based search on TRECVid AVS query sets, especially on those out-of-vocabulary queries. (3) We also investigate the impact of recent advanced text/video features on a state-of-the-art AVS model. The experimental results demonstrate that integrating the three newly introduced components into the AVS model outperforms most of the top-1 results reported on the TRECVid AVS benchmarks over the past eight years, contributing to a new state-of-the-art performance in the AVS task.

2 RELATED WORK

2.1 Ad-hoc Video Search

Ad-hoc video search, a task that has been consistently evaluated yearly in TRECVid, has its origins dating back to 2003 [48]. Given a query (i.e., a textual description of the desired search content), the search engine needs to process the query and return a ranked list of video clips [6].

From its very beginning, the mainstream approaches focus on using semantic concepts to tackle the task [50], i.e., concept-based search. Early efforts are devoted to concept bank development and ontology reasoning [25, 40, 51], and concept screening, representation and combination [1, 23, 35, 43, 55, 57]. Out-of-vocabulary (OOV) is one of the challenges in concept-based search. Existing approaches mainly address the challenge in two ways: by building a large concept bank or by having a good strategy for concept selection. The most recent methods concentrate on building a large concept bank, e.g., over 47,000 concepts [55, 57]. Besides, Nakagome et al. [58] select concepts for a query by applying ontology analysis on WordNet [39] to find hypernyms for query tokens and measure semantic similarity between concepts and the hypernyms. Huang et al. [23] also perform a series of measurements between query terms and concepts to tackle the OOV problem, such as synset similarity based on WordNet taxonomy and explicit semantic analysis based on Wikipedia. Although numerous efforts have been made [9, 23, 24, 42, 58], it still has difficulty in automatically selecting concepts for queries. Human intervention is always needed in picking concepts to boost AVS performances [55, 57]. Concept-based search performs well on those queries when their information need could be precisely identified by a list of concepts. However, the inherent problem of existing approaches is expression ambiguity. For example, it is hard to precisely convey the search ambition by using a word-only concept bank with the concepts “hold”, “hand”, and “face” for the query *Find shots of a person holding his hand to his face*, especially if a query involves prepositional words. To this end, we propose a multi-word concept bank in this paper to enhance the concept-based approach in modeling relationships between query words. Different from the existing approaches, which accumulate concepts from existing object, action, and scene classification datasets to build a concept bank, our multi-word concept bank is built based on syntax analysis of video captions. In addition to nouns and verbs, we also extract five common phrases in English (i.e., noun phrases, verb phrases, adjective phrases, prepositional phrases, and quantifier phrases) as concepts. As a result, the previous query can be identified by a verb phrase with a prepositional phrase, i.e., "hold hand" and "to face".

Embedding-based search, which encodes a user query and video clips in a high-dimensional latent space, is another mainstream approach in AVS. With more video captioning datasets (e.g., MSRVTT [65]), TGIF [33], VATEX [61]) available, embedding-based search has significantly outperformed concept-based search since 2018. Many models have been applied for AVS, including VideoStory [21], visual semantic embedding (VSE++) [17], intra-modal and inter-modal attention networks (IAN) [23], Word2VisualVec (W2VV) [13], dual encoding [14], HGR [11] and SEA [32]. The differences between these models are how they encode and represent a query text. For example, VSE++ embeds the query by a recurrent network [17], while IAN assigns different attention to the query [23]. A more complex text encoder is used in W2VV [13], which puts bag-of-words (BoW), word2vec (W2V), and word sequence altogether. The extension of W2VV (W2VV++) is the first model that significantly outperforms concept-based search on AVS [5]. [34] studies some variants of W2VV++ by reducing or replacing networks of the text encoder. Building on top of W2VV++, a recent dual encoding network [14] reports better performances by a multi-level text assembler. Three different encoders are applied, including a short-term local encoder, a word pooling encoder, and a word sequence encoder. Dual coding has some variants. For example, Damianos et al. extend it by adding an attention mechanism [19]. Besides, inspired by the recent progress of graph convolutional networks, HGR [11] encodes a query by various graphs. The most recent works, SEA [32] and LAFF [22], design query assemblers of several encoders and train the encoders in multiple spaces with multiple losses. They report state-of-the-art performances on TRECVid query sets. Nevertheless, the result of the embedding-based search is not interpretable and predictable. For instance, two queries *Find shots of a bald man* and *Find shots of a hairless man*, although they have similar information needs, the results are dramatically different.

A hybrid of concept-based search and embedding-based search has also been explored [18, 23, 41, 47, 49, 55, 58, 63]. The early studies include the fusion of VideoStory embedding and concept features [49], leading to a boost over individual models. In the recent TRECVid evaluation, lately fusing concept-based search and embedding-based search has become a norm [18, 23, 41, 47, 55, 57, 58]. Although they are shown to be complementary, the fact that both searches are produced by two individual models trained with different forms of data has tremendously increased the system complexity. To this end, the dual-task model [62] proposes to train them with the same training set of data in an end-to-end manner via dual-task learning. Specifically, it learns video-query features in a joint space and interprets the space with semantic concepts to solve the interpretation problem of the embedding-based search. Following the effort of the dual-task model, RIVRL [16] aligns a query and video clips in a feature space and a concept space, and ITV [63] proposes unlikelihood training to complement the likelihood training in [62] to have more consistent query-video embedding interpretations. As the concept and embedding-based searches are complementary, this kind of interpretable approach is a new state-of-the-art for the AVS task. In this paper, we follow the effort and propose three components to enhance an interpretable approach.

2.2 Large-scale Video-Text Datasets

In recent years, we have witnessed large-scale text-image datasets in the open domain, such as WebImageText [45] with 400M text-image pairs and LAION-5B [46] with 5B text-image pairs, have substantially improved the performances of multiple text-image tasks, e.g., text-image retrieval and image captioning. Similarly, large-scale video-text datasets are proposed to facilitate the progress of video-text tasks. For example, the HowTo100M dataset [38], which has 136M video clips covering 23k human activities and each video clip has a caption obtained from a narration, has promoted the progress of the text-based action location task and text-to-video retrieval in instruction domains such as cooking. However, as the caption is a narrative (a subtitle), it cannot be guaranteed that it is aligned with the visual content. To this end, another large-scale text-video dataset (i.e., WebVid2M [7]) is recently proposed for the open domain. It has 2.5M video clips, and each of them is associated with a manually annotated caption. However, the caption lengths and styles have big differences in this dataset. The caption lengths range from 4 to 40. Some of the captions have well-defined sentence formats, while some have less defined formats, such as keywords and disjoint sentences. Moreover, the existing large-scale video-text datasets only have one caption per video clip, where the rich semantic information of a video is overlooked. In this paper, we follow the successful attempts at generating captions for images, such as LAION-COCO[2] and ALIP [66], and construct a large-scale video-text dataset (WebVid-genCap7M), by automatically generating captions for videos. Specifically, we generate multiple captions with diverse content and different wordings for a video clip to cover the rich information in semantic and to facilitate the learning of robust alignments between a video clip and different wording sentences.

2.3 Concept Bank Construction

A good quality concept bank is effective in addressing the out-of-vocabulary (OOV) problem in concept-based search. The existing concept-based approaches usually construct a concept bank by composing various classes of multiple off-the-shelf classification datasets. For example, [23] constructs a concept bank by accumulating object classes from YFCC100M [53], action classes from UCF101 [52] and Kinetics [10], location classes from Place365 [68], and a combination of person+object+action+location classes from SIN346 [44]. However, the current way of constructing a concept bank inherently limits the ability to model a search intention with a focus on relationships and attributes. Although [55, 57] construct ATTRIBUTES300 and RELATIONSHIPS53 based on the Visual Genome dataset [26] to better model the attributes and relationships between persons and objects, the number of the concept and the fact that the concept extractors are trained on images restrict their effectivenesses on ad-hoc queries and large video corpus. Recently, the interpretable approaches [15, 16, 62, 63] built a concept bank by directly dividing a sentence into concept tokens. Although the concept bank has all the basic elements of a sentence (e.g., nouns, verbs, adjectives), they are word-only concepts. In this paper, we propose to build a multi-word concept bank based on syntax analysis of sentences to include both words and phrases. The experimental results show that the newly constructed concept bank, which includes the basic elements of a query sentence and the possible relationships of query words, has significantly improved the performance of a concept-based search, and the performance improvements are significant on the OOV queries. Moreover, the new concept bank, along with the new dataset and recent-advanced features, promotes the concept-based search as being competitive or even better than embedding-based search on some TRECVid AVS query sets.

3 IMPROVED INTERPRETABLE EMBEDDINGS

In this section, we illustrate how to plug in three proposed components to a state-of-the-art interpretable embedding model (ITV) [63]. Our proposed model is the ITV model with all three components (named improved ITV).

Given a video v and a text query q, an interpretable embedding model encodes them by a visual encoder $H(x)$ and a textual encoder $F(x)$ to a joint embedding space as $H(v) \in \mathbb{R}^d$ and $F(q) \in \mathbb{R}^d$, respectively where d is the dimension of the latent space. The video/text embedding is subsequently interpreted by a concept decoder $G(x)$ and outputs a probability vector over n concepts, e.g., $G(H(v)) = [p_1, p_2, ..., p_n]$ where p_i indicates the probability of the concept i being present in the video v. The interpretable embedding model is trained end-to-end via dual-task learning. On the one hand, visual and textual encoders are trained to ensure text-video pairs stay close in the joint space. On the other hand, the concept decoder is trained to decode concepts from a visual/textual embedding for describing the semantics in the text (video captions). In the inference stage, concept-based search aligns a query text with videos based on the similarity of $G(F(q))$ and $G(H(v))$ while embedding-based search is based on $F(q)$ and $H(v)$. Fusion search combines concept-based search and embedding-based search by a linear function.

3.1 Multi-word Concept Bank

Existing interpretable models [15, 62, 63] construct a concept bank by directly dividing text into individual word tokens without syntax analysis, and interpret embeddings with only words. Constructing a concept bank in such way neglects composition/relationship between words in a sentence, which leads to incomprehensible and imprecise interpretation. For example, interpreting the embedding of a compositional phrase (e.g., "sign language") with individual word concepts (e.g., "sign" and "language") could mislead the model understanding. Interpreting the embedding of an adjective phrase (e.g., "a blue shirt") with object/being and attribute words separately (e.g., "blue" and "shirt") will eventually lead to an imprecise interpretation. Especially, when there are multiple adjectives in a sentence, simple word tokenization will result in various mismatches of attributes and objects/beings.

In this paper, we propose to perform syntax analysis on text before building a concept bank and associate embeddings with both word and phrase concepts to provide a more comprehensive

[2] https://laion.ai/blog/laion-coco/

Table 1: Dataset statistics of existing video caption dataset on the open domain. We automatically create a new large-scale video-caption dataset named WebVid-genCap7M with multiple captions per video.

dataset	domain	caption type	#videos	#captions	#avg token	#cap/video
MSRVTT [65]	open	manually annotated	10K	100K	9.28	10
TGIF [33]	open	manually annotated	100K	128k	11.28	1
VATEX [61]	open	manually annotated	34.9k	349k	15.23	10
HowTo100M [38]	instruction	subtitles	136M	136M	4.16	1
WebVid2M [7]	open	manually annotated	2.5M	2.5M	12	1
WebVid-genCap7M	open	automatically generated	1.44M	7.1M	9.94	5

Algorithm 1: Multi-word concept bank construction

```
Input  : A sentence corpus C
Output : A concept bank CB
Initialize a concept counter counter;
Set targetedPhraseTypes=[noun phrases (NP), verb phrases
  (VP), adjective phrases (ADJP), prepositional phrases (PP),
  and quantifier phrases (QP)];
Concept Extraction;
for each sentence s in C do
    Initialize conceptList =[];
    Extract a parse tree of the sentence Ptree = Parser(s);
    for each node in the Ptree do
        if current node is leaf then // single word
            if node['attr'] is verbs or nouns then
                conceptList← node['word'];
            end
        else
            if length(node['word']) is 2/3/4 then
             // two/three/four-word phrases
                if node['attr'] is in wantedPhraseTypes then
                    conceptList← node['word'];
                end
            end
        end
    end
    counter.update(conceptList);
end
Concept Bank Construction;
Initialize CB=[];
for each concept in counter do
    if counter[concept] > 20 then
        CB←concept;
    end
end
```

and precise interpretation. The process of building a multi-word concept bank is illustrated in Algorithm 1. It has two stages: concept extraction and concept bank construction. For each sentence in the corpus, we extract word and phrase concepts. Specifically, given a sentence, we first perform syntax analysis and obtain its parse tree. Next, the algorithm goes through each node of the parse tree to extract nouns, verbs, and five main types of phrases in English: noun phrases, verb phrases, adjective phrases, prepositional phrases, and quantifier phrases. After concept extraction, we count the frequency of a concept that appears in the corpus, and those concept words or phrases that appear more than 20 times are used to build the concept bank.

3.2 Large Video-GenText Dataset for Pre-training

To learn effective and representative cross-modal interpretable embeddings, having sufficient annotation data is essential. Furthermore, an image is worth a thousand words, and a video is a series of images. Having multiple captions for a video is important to learn robust alignments between a video and sentences. However, the existing video caption datasets are either small size (e.g., MSRVTT [65], TGIF [33], VATEX [61]) or have a small number of captions per video (e.g., HowTo100M [38], WebVid2M [7]) as shown in Table 1. Generating pseudo labels for unlabeled data in supervised learning has shown effectiveness in improving the performance [27], and there are some successful attempts at generating synthetic captions for image dataset [66] to improve text-image retrieval. Following these efforts and capitalizing on the recent progress of pre-trained generative models, we explore generating captions for videos to supervise the training of cross-modal interpretable embeddings.

In this paper, we generate synthetic captions for videos[3] in the WebVid2M dataset [7] to pre-train the interpretable embedding model. Specifically, for each video, we extract around five frames, and the duration of two adjacent frames is around 3.6 seconds to ensure visual difference. For each frame, we generate a caption using an image captioning model [29]. Eventually, a video is associated with multiple generated captions for cross-modal interpretable embedding learning. In total, we generate seven million captions for about 1.44 million videos. We name this dataset WebVid-genCap7M. As shown in Table 1, compared to the existing dataset, WebVid-genCap7M has a larger scale and a larger number of captions per video than other huge datasets. By having denser frame sampling or having more captions for a frame, the number of captions can become larger. Figure 1 shows some example video-caption pairs in WebVid-genCap7M along with the original video captions in WebVid2M [7] for comparisons. The captions of WebVid2M [7] have

[3]We only manage to download 1.44M videos based on the official provided video URL.

Original caption	"Santa claus in red costume talking, smiling and shake head on green chroma key background. christmas and new year celebration concept"	"Pov footage of boat trip on rhine river. picturesque landscape of mosel region, germany. summer nature background, timelapse. 4k, ultra high definition, ultra hd, uhd, 2160p, 3840 x 2160"	"Cali, colombia - july 03, 2018: little girl jumps in park in park in cali, colombia; south america"
Generated caption	1. A man in a santa suit with a surprised look on his face 2. A man dressed as santa claus on a green screen 3. A man dressed as santa claus laughing and laughing 4. A man in a santa suit is making a funny face 5. A man dressed in a santa claus costume	1. A river running through a lush green hillside 2. A body of water surrounded by mountains and trees	1. A young girl drinking from a pink cup 2. A little girl standing in the dirt with a pink object in her mouth 3. A little girl standing in front of a bench 4. A little girl standing in the dirt with a pink cup 5. A little girl standing on top of a dirt field

Figure 1: Example video-GenCaption pairs from the WebVid-genCap7M dataset along with the original captions in WebVid2M [7].

large differences in length and format. For example, the lengths of the original caption in the second and third example in Figure 1 are 29 and 17, respectively. They also have different formats. For example, the first original caption is in sentence format, while the third is in keyword format. In contrast, our generated captions are similar in length and have well-defined sentence structure. Moreover, they are diverse in wording expressions and mention extra details that are correct for the video but not mentioned in the original caption, such as "a young girl drinking from a pink cup" in the third example.

3.3 Features Enhancement

As transformers show great effectiveness in multiple cross-modal tasks, we enhance both the textual and visual encoders of a state-of-the-art interpretable embedding model [63] with pre-trained transformers. Specifically, the textual encoder in [63] consists of bag-of-words (BoW), W2V, and GRU, i.e., $F(x) = BN(FC([\text{BoW, W2V, biGRU}]))$, where BN and FC are a batch normalization layer and a fully-connected layer. The W2V is pre-trained on English tags of 30 million Flickr images [67]. We replace the textual encoder with three recent advanced pre-trained text-visual transformers, i.e., CLIP [45], BLIP-2 [28], and imagebind [20], i.e., $\hat{F}(x) = BN(FC([\text{CLIP, BLIP-2, imagebind}]))$. Their weights are frozen during training. Similarly, we add these three pre-trained transformers to the visual encoder in [63] and also freeze their weights in the training. In other words, the original encoder $H(x) = BN(FC([\text{CNNs,biGRU, biGRU-CNN, swinTrans, SlowFast}]))$ is changed to $\hat{H}(x) = BN(FC([\text{CNNs, biGRU, biGRU-CNN, swinTrans, SlowFast, CLIP, BLIP-2, imagebind}]))$. The weights of CNNs, swinTrans, and SlowFast are frozen as the same as in [63].

4 EXPERIMENTS

This section first evaluates the newly introduced components on TRECVid AVS datasets [6, 8] through ablation studies. We also compare the proposed methods with the state-of-the-art on the MSRVTT [65] and TRECVid AVS datasets.

4.1 Experiment Setting

4.1.1 Datasets. In the pre-training stage, we use the WebVid-GenCap7M for training, and WebVid2M [7] val set for validation. Following the setting of the existing AVS approaches [31, 62, 63], we fine-tune the interpretable embedding model on the combination of TGIF [33], MSRVTT [65], and VATEX [61] datasets, validate it on a TRECVid VTT dataset (i.e., tv2016train) [48] and test it on TRECVid AVS datasets (i.e., IACC.3 [6], V3C1 [8], and V3C2 [8]). IACC.3 dataset has 334k video clips associated with 90 AVS queries across three years. V3C1 and V3C2 have 1 million and 1.4 million video clips, respectively, where the former is linked with 70 AVS queries used in 2019-2021, and the latter consists of 50 AVS queries used in 2022-2023.

4.1.2 Evaluation Metric. We follow the AVS standard [48] to report xinfAP with a search length of 1000 on TRECVid AVS datasets. For the text-to-video retrieval on the MSRVTT dataset, we report R@1,5,10, MedR, and mAP.

Table 2: AVS Performance comparison between word-only and multi-word concept banks, and with and without WebVid-genCap7M in pre-training (PreT)

	Concept-based search			Embedding-based search			Fusion search		
concept bank type	word	word	multi-word	word	word	multi-word	word	word	multi-word
with/without pre-training	w/o PreT	w/ PreT	w/o PreT	w/o PreT	w/ PreT	w/o PreT	w/o PreT	w/ PreT	w/o PreT
tv16	0.184	0.171	**0.197**	0.187	**0.208**	0.179	0.211	**0.212**	0.209
tv17	0.230	0.240	**0.288**	0.279	**0.290**	0.283	0.292	0.302	**0.325**
tv18	0.135	0.127	**0.162**	**0.140**	0.139	0.134	**0.170**	0.158	0.169
tv19	0.166	0.167	**0.197**	0.201	**0.205**	0.203	**0.227**	0.216	0.222
tv20	**0.292**	0.233	0.285	0.307	0.312	**0.319**	0.345	0.321	**0.346**
tv21	0.246	0.262	**0.267**	0.294	**0.295**	0.284	**0.318**	**0.318**	0.308
tv22	0.115	0.090	**0.116**	**0.135**	0.123	0.131	**0.150**	0.119	0.149
tv23	0.124	0.089	**0.186**	0.151	**0.160**	0.153	0.167	0.147	**0.195**
mean	0.186	0.172	**0.212**	0.212	**0.216**	0.211	0.235	0.224	**0.240**

Table 3: Performance comparison on feature enhancement. $F(x)$ and $H(x)$ are the baseline textual end visual encoders, while $\hat{F}(x)$ and $\hat{H}(x)$ are with enhanced textual and visual features.

		IACC.3			V3C1			V3C2		
textual encoder	visual encoder	tv16 (30)	tv17 (30)	tv18 (30)	tv19 (30)	tv20 (20)	tv21 (20)	tv22 (30)	tv23 (20)	mean
$F(x)$	$H(x)$	0.211	0.292	**0.170**	0.227	0.345	0.318	0.150	0.167	0.235
$\hat{F}(x)$	$H(x)$	0.216	0.288	0.148	0.204	0.325	0.307	0.150	0.173	0.226
$\hat{F}(x)$	$\hat{H}(x)$	0.249	0.279	0.168	0.243	0.360	0.364	**0.215**	**0.250**	0.266
$F(x)$	$\hat{H}(x)$	**0.254**	**0.318**	0.162	**0.254**	**0.364**	**0.368**	0.179	0.241	0.**268**

4.1.3 Implementation Details. We implement our proposed methods based on the publicly available code provided by the interpretable embedding model (ITV) [63] and use the same parameter setting as ITV. We follow ITV to use cosine similarity to measure the alignment between the query and videos and set equal weight for the concept-based search and embedding-based search in the fusion. For the concept bank extraction, we use Stanford coreNLP parser[4] to obtain the parse tree of a sentence. For the image caption generation, we use the base and large BLIP models [29] trained on the MSCOCO dataset [12].

4.2 Ablation Studies

The section studies the impact of the proposed three components, and the original interpretable embedding model [63] is used as a baseline.

4.2.1 Word-only versus Multi-word Concept Banks. Figure 2 visualizes the phrases and their frequencies in the multi-word concept bank built on a caption corpus. The corpus contains all the video captions of TGIF [33], MSRVTT [65], and VATEX [61] datasets. The concept bank has 14,528 concepts, 9,465 of which are phrases. 62% of phrases appear between 20 to 50 times, and 18% appear more than 100 times in the training corpus. As shown in Figure 2, the concept bank manages to contain five main types of phrases, including noun phrases such as *man and woman*, verb phrases such as *sit down*, adjective phrases such as *young man*, prepositional phrases such as *on floor* and quantifier phrases such as *two man*.

[4]https://nlp.stanford.edu/software/srparser.html

We compare the AVS performances with two different concept banks in interpreting embeddings without pre-training. Table 2 contrasts the retrieval results on eight query sets across three search modes: concept-based search, embedding-based search, and the fusion of them. Our proposed multi-word concept bank significantly outperforms the word-only concept bank consistently on the concept-based search across most query sets and boosts the concept-based search to be competitive with the embedding-based search interpreted by a word-only concept bank. About 65% of queries get improved on the concept-based search, and about 57% of them are bad-performing queries (i.e., xinfAP < 0.1) that suffer from out-of-vocabulary problems. The performance improvement is mainly attributed to a better capability of modeling the relationships between query words. For example, for query-535 *Find shots of a person standing in front of a brick building or wall*, it is almost impossible to use a list of word-only concepts to interpret the position of the person and building/wall. However, with the addition of the prepositional phrase *in front brick wall*, the position relation can be interpreted properly, and the retrieval performance increases by six times. As embedding-based search is good at modeling the relationships of query words, the addition of the phrase concepts does not bring a significant improvement to embedding-based search, and the average retrieval performances of the two concept banks are almost the same. Overall, with the improvement in concept-based search, the multi-word concept bank has a slightly better average xinfAP than the word-only on fusion search.

4.2.2 Impact of Pre-training on Video-GenText Pairs. We evaluate the impact of the proposed large-scale video-text dataset in

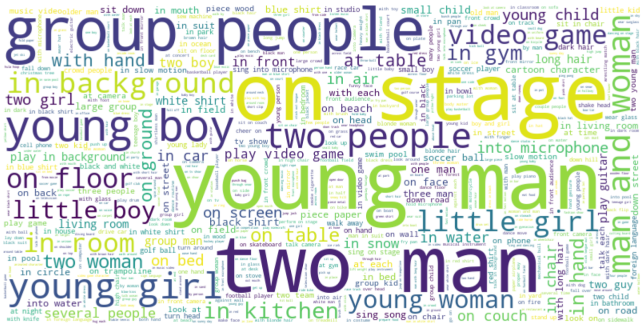


Figure 2: Cloud figure of phrases in the multi-word concept bank.

Table 4: Comparison on the MSRVTT dataset. All performances are cited directly from the published results. Except CLIP4CLIP is rerun and tested on the same split of the MSRVTT dataset. The unavailable results are marked with /.

	Text-to-Video Retrieval				
	R@1	R@5	R@10	Med r	mAP
W2VV++ [31]	11.1	29.6	40.5	18	20.6
Dual encoding [14]	11.1	29.4	40.3	19	20.5
Hybrid space [15]	11.6	30.3	41.3	17	21.2
Dual-task	12.2	31.7	42.9	16	22.3
SEA [32]	13.1	33.4	45.0	14	23.3
RIVRL [16]	13.0	33.4	44.8	14	23.2
LAFF [22]	16.0	39.5	51.4	10	27.6
ITV [63]	13.2	33.7	45.0	14	23.4
CLIP4CLIP [36]	28.6	53.2	63.8	5	/
improved ITV	**29.3**	**54.4**	**65.0**	**4**	**41.2**

pre-training interpretable embeddings. Table 2 compares the AVS performance of interpretable embedding with and without pre-training on the WebVid-genCap7M dataset with the word-only concept bank. The pre-trained model has better performances on six out of eight query sets and obtains the highest overall performance on embedding-based search. Specifically, taking advantage of the pre-training, over 50% of queries have higher xinfAP on embedding-based search. However, on the concept-based search, the pre-training degrades the xinfAPs on five out of eight query sets, and there are considerable drops on the tv22 and tv23 query sets. As observed, these are mainly attributed to the ground truth problem. For example, about 60% of the video clips retrieved by the pre-trained model on concept-based search on the tv22 query set are newly found and not judged in the evaluation in that year. When computing the xinfAP metric, the unjudged video shots are regarded as wrong results. We compute a xinfAP with a shorter search length (i.e., 10) where there is a smaller percentage of unjudged video clips, the pre-trained model exceeds the model without pre-training with xinfAP@10=0.281 to xinfAP@10=0.278 on the tv22.

4.2.3 Impact of Advanced Textual and Visual Features. In this section, we investigate whether the integration of recent advanced textual/visual features to the interpretable embedding model (ITV) [63] improves the retrieval performance. The experiment is conducted on TRECVid AVS datasets with eight query sets. The original setting of the ITV is used as a baseline. We change the textual encoder from $F(x)$ to $\hat{F}(x)$ and the visual encoder from $H(x)$ to $\hat{H}(x)$ in Section 3.3 to include recent advanced features, i.e., CLIP [45], BLIP-2 [28], and imagebind [20],

Table 3 compares the AVS performances with different combinations of encoders. On the text side, the transformer-based pre-trained models are not comparable to the traditional encoders on most AVS query sets and degrade the mean performance of the baseline by 4%. In contrast, except for the tv17 query set, the addition of the advanced visual features does boost the retrieval performances consistently by a large margin. It also beats the enhanced features on both text and video sides on most query sets.

4.3 Text-to-video Retrieval on MSRVTT Dataset

We compare the proposed model with other existing methods on the MSRVTT dataset [65]. The proposed model is with the multi-word concept bank, pre-trained on the WebVid-genCap7M dataset, and with enhanced features on both text and video sides. Table 5 reports the retrieval results on the official split, which has 6513, 497, and 2990 videos for training, validation, and testing, respectively. With all the three proposed components, the improved ITV manages to exceed other approaches significantly. Especially, it outperforms the original ITV [64] more than two times on R@1 and has higher mAP on 64.2% test queries. Furthermore, compared to CLIP4CLIP [36], which uses the pre-trained weights of CLIP and fine-tunes on MSRVTT, the improved ITV also shows better performances on all evaluated metrics.

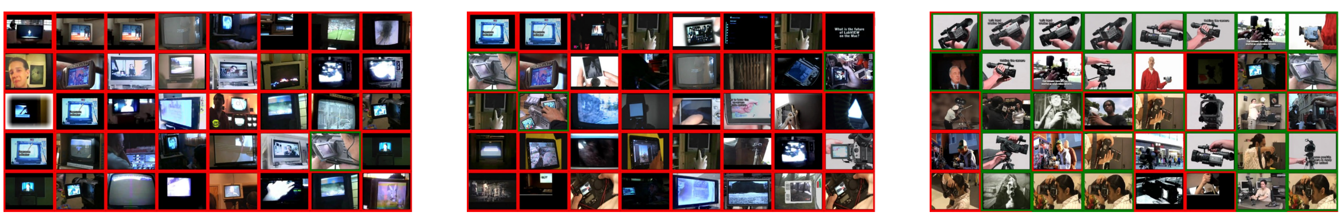
4.4 AVS Comparison to the State-of-the-art

Table 5 compares our proposed method with widely used pre-trained models, existing AVS approaches, and the top-1 performances reported by TRECVid in each year of evaluation. For the recent approaches, LAFF [22] and RIVRL [16], we reproduce their results using the same setting as ours for fair comparisons (e.g., the same training sets and visual features) based on their published codes and report the results across eight years of query sets (tv16 to tv23).

As shown in Table 5, our proposed improved ITV significantly outperforms the pre-trained models on large image-text datasets and all the existing AVS approaches by a large margin. Specifically, compared to the most recent AVS approaches, LAFF and RIVRL, our model has better performances on 74.3% and 80% of queries, respectively. It also manages to improve 31.5% of queries that LAFF and RIVRL both have xinfAP lower than 0.1 to over 0.1. Figure 3 shows an example contrasting LAFF and RIVRL on the query-554. As a TV or movie camera is not frequently seen in video caption datasets, LAFF and RIVRL find video shots of TV instead. With the large-scale pre-trained WebVid-genCap7M, the improved ITV outperforms them significantly, even though some correct results are not annotated(judged) in the ground truth. Furthermore, the improved ITV also outperforms or is competitive with the top-1 solution on most query sets even though the top-performance teams in TRECVid AVS evaluation usually fuse multiple rank lists from

Table 5: Performance comparison across eight years of TRECVid AVS datasets. The number inside parentheses indicates the number of queries evaluated that year. The reproduced results are marked with *

Datasets	IACC.3			V3C1			V3C2	
Query sets	tv16 (30)	tv17 (30)	tv18 (30)	tv19 (30)	tv20 (20)	tv21 (20)	tv22 (30)	tv23 (20)
TRECVid top result:								
top-1	0.054	0.206	0.121	0.163	**0.359**	0.355	**0.282**	0.292
Pre-trained models:								
CLIP [45]	0.182	0.217	0.089	0.117	0.128	0.178	0.124	0.109
BLIP-2 [28]	0.213	0.226	0.168	0.199	0.222	0.273	0.164	0.203
CLIP4CLIP [36]	0.182	0.217	0.089	0.133	0.149	0.188	0.121	0.109
AVS approaches:								
ConBank	/	0.159 [54]	0.060 [58]	/	/	/	/	/
ConBank (manual)	0.177 [56]	0.216 [54]	0.106 [58]	0.114 [55]	0.183 [57]	/	/	/
W2VV++ [31]	0.150	0.207	0.099	0.146	0.199	/	/	/
Dual coding [14]	0.160	0.232	0.120	0.163	0.208	/	/	/
Dual-task [62]	0.184	0.252	0.120	0.189	0.229	0.193	/	/
HGR [11]	/	/	/	0.142	0.301	/	/	/
SEA [32]	0.164	0.228	0.125	0.167	0.186	/	/	/
Hybrid space [15]	0.157	0.236	0.128	0.170	0.191	0.162	/	/
LAFF* [22]	0.188	0.261	0.152	0.215	0.299	0.300	0.178	0.172
RIVRL* [16]	0.159	0.231	0.131	0.197	0.278	0.254	0.179	0.177
$\text{ITV}_{concept}$ [63]	0.184	0.230	0.135	0.166	0.292	0.246	0.115	0.124
$\text{ITV}_{embedding}$ [63]	0.187	0.279	0.140	0.201	0.307	0.294	0.135	0.151
ITV_{fusion} [63]	0.211	0.292	**0.170**	0.227	0.345	0.318	0.150	0.167
The proposed models:								
improved $\text{ITV}_{concept}$	0.252	0.310	0.127	0.161	0.245	0.295	0.164	0.280
improved $\text{ITV}_{embedding}$	0.233	0.296	0.167	0.237	0.334	0.309	0.198	0.241
improved ITV_{fusion}	**0.280**	**0.349**	0.165	**0.242**	0.352	**0.365**	0.235	**0.295**



(a) LAFF (xinfAP=0.039) (b) RIVRL (xinfAP=0.011) (c) Improved ITV (xinfAP=0.352)

Figure 3: Comparison with the state-of-the-art approaches LAFF and RIVRL on query-554 *Find shots of a person holding or operating a tv or movie camera.*

various models, such as the top-1 solution on the tv22 query set fusing the results of more than 100 rank lists from five models.

The performance improvements over the original ITV are consistent across concept-based, embedding-based, and fusion searches. There are 139, 136, and 134 out of 210 queries having higher performances on the three search modes, respectively. The number of queries whose xinfAP are less than 0.1 also decreases by 13%, 31%, and 27% on three search modes, respectively. This is mainly due to the effectiveness of the three proposed components in addressing the out-of-vocabulary problem suffered by these queries. The three proposed components are complementary as their combination outperforms the performances of all three individual components as shown in ablation studies.

5 CONCLUSION

In this paper, we study three components to address the small-size dataset and out-of-vocabulary problems on the AVS task. The multi-word concept bank boosts the performance of out-of-vocabulary queries by enhancing the capability of concept-based search on modeling relationships between query words. The newly constructed video-GenText dataset manages to improve the embedding-based search by having more training instances on unseen queries. The recent-advanced visual features manage to increase the retrieval performance, while the advanced textual features are not competitive with the traditional features on the AVS task. The three introduced elements are shown to be complementary, and their combination significantly increases the retrieval performances of an interpretable embedding model and outperforms the state-of-the-art AVS approaches on both small (i.e., MSRVTT) and large datasets (i.e., TRECVid AVS datasets).

REFERENCES

[1] Konstantinos Avgerinakis, Anastasia Moumtzidou, Damianos Galanopoulos, Georgios Orfanidis, Stelios Andreadis, Foteini Markatopoulou, Elissavet Batziou, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. 2018. ITI-CERTH participation in TRECVid 2018. In *Proceedings of the TRECVid 2018 Workshop*. 1–13.

[2] George Awad, Asad A. Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, Gareth J. F. Jones, and Georges Quénot. 2021. Evaluating Multiple Video Understanding and Retrieval Tasks at TRECVid 2021. In *Proceedings of TRECVid 2021*. 1–55.

[3] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. 2020. TRECVid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVid 2020*. 1–55.

[4] George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Eliot Godard, Lukas Diduch, Yvette Graham, and Georges Quénot. 2023. TRECVID 2023 - A series of evaluation tracks in video understanding. In *Proceedings of TRECVID 2023*. NIST, USA, 1–23.

[5] George Awad, Asad Gov, Asad Butt, Keith Curtis, Yooyoung Lee, yooyoung@nist Gov, Jonathan Fiscus, David Joy, Andrew Delgado, Alan Smeaton, Yvette Graham, Wessel Kraaij, Georges Quenot, Joao Magalhaes, and Saverio Blasi. 2018. TRECVid 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. In *Proceedings of the TRECVid 2018 Workshop*.

[6] George Awad, Fiscus Jonathan, Joy David, Michel Martial, Smeaton Alan, Kraaij Wessel, Quenot Georges, Eskevich Maria, Aly Robin, Ordelman Roeland, Jones Gareth, Huet Benoit, and LarsonMartha. 2016. TRECVid 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *Proceedings of the TRECVid 2016 Workshop*. 1–54.

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision*. 1–15.

[8] Fabian Berns, Luca Rossetto, Klaus Schoeffmann, Christian Beecks, and George Awad. 2019. V3C1 Dataset: An Evaluation of Content Characteristics. In *Proceedings of the International Conference on Multimedia Retrieval*. 334–338.

[9] Maaike H. T. De Boer, Yi-Jie Lu, Hao Zhang, Klamer Schutte, Chong-Wah Ngo, and Wessel Kraaij. 2017. Semantic Reasoning in Zero Example Video Event Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 4 (2017), 1–17.

[10] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. *ArXiv* abs/1808.01340 (2018), 1–6.

[11] S. Chen, Y. Zhao, Q. Jin, and Q. Wu. 2020. Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10635–10644.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR* abs/1504.00325 (2015), 1–7.

[13] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. 2018. Predicting Visual Features From Text for Image and Video Caption Retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.

[14] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual Encoding for Zero-Example Video Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9346–9355.

[15] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual Encoding for Video Retrieval by Text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–17.

[16] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. 2022. Reading-Strategy Inspired Visual Representation Learning for Text-to-Video Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 32 (2022), 5680–5694.

[17] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidlere. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference*. 1–13.

[18] Danny Francis, Phuong Anh Nguyen, Benoit Huet, and Chong-Wah Ngo. 2019. EURECOM at TRECVid AVS 2019. In *Proceedings of the TRECVid 2019 Workshop*.

[19] Damianos Galanopoulos and Vasileios Mezaris. 2020. Attention Mechanisms, Signal Encodings and Fusion Strategies for Improved Ad-Hoc Video Search with Dual Encoding Networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 336—-340.

[20] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *CVPR*. 1–11.

[21] Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek. 2014. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *Proceedings of the ACM Conference on Multimedia*. 17–26.

[22] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. 2022. Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval. In *European Conference on Computer Vision*. Springer Nature Switzerland, Cham, 444–461.

[23] Po-Yao Huang, Junwei Liang, Vaibhav, Xiaojun Chang, and Alexander Hauptmann. 2018. Informedia @ TRECVid 2018: Ad-hoc Video Search with Discrete and Continuous Representations. In *Proceedings of the TRECVid 2018 Workshop*. 1–10.

[24] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander Hauptmann. 2015. Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*.

[25] Yu-Gang Jiang, Jun Yang, Chong-Wah Ngo, and Alexander Hauptmann. 2010. Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study. *IEEE Transactions on Multimedia* 12 (2010), 42–53.

[26] Ranjay Krishna, Yuke Zhu, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Michael S. Bernstein, and Li Fei-Fei. [n. d.]. ([n. d.]).

[27] Dong-Hyun Lee. 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)* (07 2013), 1–6.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ArXiv* abs/2301.12597 (2023), 1–13.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*. 1–12.

[30] Xirong Li, Jianfeng Dong, Chaoxi Xu, Jing Cao, Xun Wang, and Gang Yang. 2018. Renmin University of China and Zhejiang Gongshang University at TRECVid 2018: Deep Cross-Modal Embeddings for Video-Text Retrieval. In *Proceedings of the TRECVid 2018 Workshop*. 1–6.

[31] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2VV++: Fully Deep Learning for Ad-hoc Video Search. In *Proceedings of the ACM International Conference on Multimedia*. 1786–1794.

[32] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. 2021. SEA: Sentence Encoder Assembly for Video Retrieval by Textual Queries. *IEEE Transactions on Multimedia* (2021), 4351–4362.

[33] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A New Dataset and Benchmark on Animated GIF Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4641–4650.

[34] Jakub Lokoć, Tomáš Soućek, Patrik Veselý, František Mejzlík, Jiaqi Ji, Chaoxi Xu, and Xirong Li. 2020. A W2VV++ Case Study with Automated and Interactive Text-to-Video Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2553—-2561.

[35] Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. 2016. Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 127–134.

[36] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *arXiv preprint arXiv:2104.08860* (2021), 1–14.

[37] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2017. Query and Keyframe Representations for Ad-Hoc Video Search. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 407–411.

[38] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*. 1–11.

[39] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[40] Milind Naphade, J.R. Smith, Jelena Tesic, S. Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. 2006. Large-Scale Concept Ontology for Multimedia. *IEEE Transactions on Multimedia* 13 (2006), 86–91.

[41] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. 2017. VIREO @ TRECVid 2017: Video-to-Text, Ad-hoc Video Search and Video Hyperlinking. In *Proceedings of the TRECVid 2017 Workshop*.

[42] Phuong Anh Nguyen, Jiaxin Wu, Chong-Wah Ngo, Francis Danny, and Huet Benoit. 2019. VIREO-EURECOM @ TRECVid 2019: Ad-hoc Video Search. In *Proceedings of the TRECVid 2019 Workshop*. 1–8.

[43] Vinh-Tiep Nguyen, Duy-Dinh Le, Benjamin Renoust, Thanh Duc Ngo, Minh-Triet Tran, Duc Anh Duong, and Shinichi Satoh. 2016. NII-HITACHI-UIT at TRECVid 2016 Ad-hoc Video Search: Enriching Semantic Features using Multiple Neural Networks. In *Proceedings of the TRECVid 2016 Workshop*. 1–4.

[44] Paul Over, Jon Fiscus, Gregory Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. 2014. TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics.

[45] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*. 8748–8763.

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems*. 1–50.

[47] Kimiaki Shirahama, Daichi Sakurai, Takashi Matsubara, and Kuniaki Uehara. 2019. Kindai University and Kobe University at TRECVid 2019 AVS Task. In *Proceedings of the TRECVid 2019 Workshop*.

[48] Alan F. Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation Campaigns and TRECVid. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*. 321–330.

[49] Cees G. M. Snoek, Xirong Li, Chaoxi Xu, and Dennis C. Koelma. 2017. University of Amsterdam and Renmin University at TRECVid 2017: Searching Video, Detecting Events and Describing Video. In *Proceedings of the TRECVid 2017 Workshop*. 1–6.

[50] Cees G. M. Snoek and Marcel Worring. 2009. Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval* 2 (2009), 215–322.

[51] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. 2006. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*. 421–430.

[52] Khurram Soomro, Amir Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR* (12 2012).

[53] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[54] Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. 2017. Waseda Meisei at TRECVid 2017: Ad-hoc Video Search. In *Proceedings of the TRECVid 2017 Workshop*. 1–8.

[55] Kazuya Ueki, Takayuki Hori, and Tetsunori Kobayashi. 2019. Waseda Meisei SoftBank at TRECVid 2019: Ad-hoc Video Search. In *Proceedings of the TRECVid 2019 Workshop*. 1–7.

[56] Kazuya Ueki, Kotaro Kikuchi, Susumu Saito, and Tetsunori Kobayashi. 2016. Waseda at TRECVid 2016: Ad-hoc Video Search. In *Proceedings of the TRECVid 2016 Workshop*. 1–5.

[57] Kazuya Ueki, Ryo Mutou, Takayuki Hori, Yongbeom Kim, and Yuma Suzuki. 2020. Waseda Meisei SoftBank at TRECVid 2020: Ad-hoc Video Search. In *Proceedings of the TRECVid 2020 Workshop*. 1–7.

[58] Kazuya Ueki, Yu Nakagome, Koji Hirakawa, Kotaro Kikuchi, Yoshihiko Hayashi, Tetsuji Ogawa, and Tetsunori Kobayashi. 2018. Waseda Meisei at TRECVid 2018: Ad-hoc Video Search. In *Proceedings of the TRECVid 2018 Workshop*. 1–7.

[59] Kazuya Ueki, Yuma Suzuki, Hiroki Takushima, Hideaki Okamoto, Hayato Tanoue, , and Takayuki Hori. 2022. Waseda Meisei SoftBank at TRECVID 2022. In *Proceedings of the TRECVid 2022 Workshop*. 1–5.

[60] Kazuya Ueki, Yuma Suzuki, Hiroki Takushima, Haruki Sato, Takumi Takada, Hideaki Okamoto, Hayato Tanoue, Takayuki Hori, and Aiswariya Manoj Kumar3. 2023. Waseda Meisei SoftBank at TRECVID 2023. In *Proceedings of the TRECVid 2023 Workshop*. 1–8.

[61] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *Proceedings of the IEEE International Conference on Computer Vision*. 4580–4590.

[62] Jiaxin Wu and Chong-Wah Ngo. 2020. Interpretable Embedding for Ad-Hoc Video Search. In *Proceedings of the ACM International Conference on Multimedia*. 3357–3366.

[63] Jiaxin Wu, Chong-Wah Ngo, Wing-Kwong Chan, and Zhijian Hou. 2023. (Un)likelihood Training for Interpretable Embedding. *ACM Transactions on Information Systems* 42, 3 (2023), 1–26.

[64] Jiaxin Wu, Phuong Anh Nguyen, Zhixin Ma, and Chong-Wah Ngo. 2021. SQL-Like Interpretable Interactive Video Search. In *International Conference on Multi-Media Modeling*. 391–397.

[65] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[66] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. ALIP: Adaptive Language-Image Pre-training with Synthetic Caption. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–10.

[67] P. Young, M. Hodosh A. Lai, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* (2014), 67–78.

[68] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition Using Places Database. In *Proceedings of the International Conference on Neural Information*. 487–495.