On Semantic Similarity in Video Retrieval

Wray, M.; Doughty, H.; Damen, D.

Link to publication

Citation for published version (APA):
Wray, M., Doughty, H., & Damen, D. (2021). On Semantic Similarity in Video Retrieval. In *Proceedings, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition: virtual, 9-25 June 2021* (pp. 3649-3659). (CVPR). Conference Publishing Services, IEEE Computer Society. https://doi.org/10.48550/arXiv.2103.10095, https://doi.org/10.1109/CVPR46437.2021.00365

On Semantic Similarity in Video Retrieval

Michael Wray   Hazel Doughty*   Dima Damen

Department of Computer Science, University of Bristol, UK

Abstract

Current video retrieval efforts all found their evaluation on an instance-based assumption, that only a single caption is relevant to a query video and vice versa. We demonstrate that this assumption results in performance comparisons often not indicative of models' retrieval capabilities. We propose a move to semantic similarity video retrieval, where (i) multiple videos/captions can be deemed equally relevant, and their relative ranking does not affect a method's reported performance and (ii) retrieved videos/captions are ranked by their similarity to a query. We propose several proxies to estimate semantic similarities in large-scale retrieval datasets, without additional annotations. Our analysis is performed on three commonly used video retrieval datasets (MSR-VTT, YouCook2 and EPIC-KITCHENS).

1. Introduction

Video understanding approaches which incorporate language have demonstrated success in multiple tasks including captioning [63, 68], video question answering [70, 74] and navigation [5, 28]. Using language to search for videos has also become a popular research problem, known as video retrieval. Methods learn an underlying multi-modal embedding space to relate videos and captions. Along with large-scale datasets [18, 42, 58, 66, 73], several video retrieval benchmarks and challenges [2, 72] compare state-of-the-art, as methods inch to improve evaluation metrics such as Recall@K and Median Rank.

In this paper, we question the base assumption in all these datasets and benchmarks—that the only relevant video to a caption is the one collected with that caption. We offer the first critical analysis of this assumption, proposing semantic similarity relevance, for both evaluation and training. Our effort is inspired by works that question assumptions and biases in other research problems such as VQA [27, 71], metric learning [47], moment retrieval [51], action localisation [4] and action recognition [16, 35, 46].

As shown in Fig. 1, current approaches target instance-based retrieval—that is, given a query caption such as *"A man doing an origami tutorial"*, only one origami video is

---

*Now at University of Amsterdam.

Query: "A man doing an origami tutorial"

Figure 1: All current video retrieval works treat caption collected for a certain video as relevant, even when other videos are equally relevant to a query text. This makes the evaluation of popular datasets ad hoc at times. We propose to use continuous similarity, allowing multiple videos to be treated as equally relevant. Ex. from MSR-VTT [66].

considered as the correct video to retrieve. In fact, many videos within the dataset can be similar to the point of being identical. The order in which such videos are retrieved should not affect the evaluation of a method. Instead, we propose utilising semantic similarity between videos and captions, where we assign a similarity score between items of differing modalities. This allows multiple videos to be considered relevant to a caption and provides a way of ranking videos from most to least similar.

Our contributions can be summarised: (i) We expose the shortcoming of instance retrieval in current video retrieval benchmarks and evaluation protocols. (ii) We propose video retrieval with semantic similarity, both for evaluation and training, where videos are ranked by their similarity to a caption, allowing multiple videos to be considered relevant and vice-versa. (iii) Avoiding large annotation effort, we propose several proxies to predict semantic similarities, using caption-to-caption matching. (iv) We analyse three benchmark datasets, using our semantic similarity proxies, noting their impact on current baselines and evaluations.

2. Related Work

We review image retrieval works that use semantic knowledge then discuss current approaches to video retrieval.

2.1. Semantic Image Retrieval

While most works focus on instance-based retrieval, a few works have explored semantic-based image retrieval.

Early works attempted to manually annotate small-scale datasets with semantic knowledge. Oliva *et al.* [49] defined

three axes of semantic relevance (e.g. artificial vs natural) in order to relate images. Using categories instead, Ojala *et al.* [48] asked annotators to split images within a dataset into discrete categories. They then considered all images within the same category as relevant.

In their investigative work, Enser *et al.* [23] showcase that semantic relevance cannot be gleaned from images alone, as it requires the knowledge of places, societal class *etc.* Barz and Denzler [9] draw a similar conclusion that visual similarity does not imply semantic similarity and so project images into a class embedding space learned from WordNet [44]. Chen *et al.* [14] instead learn two spaces, one for images and one for text, with the notion that features in either space should be consistent if they are semantically relevant. Gordo and Larlus [26] train their model for image-to-image retrieval with the notion of semantic relevance. By learning an embedding using semantic proxies (METEOR [8], tf-idf and Word2Vec [43]) defined between image captions, they show that semantic knowledge improves retrieval performance. Concurrent with out work, Chun *et al.* [17] highlight the issue of instance based evaluation for cross-modal retrieval in images. They propose using R-Precision as an evaluation metric incorporating further plausible matches via class knowledge. However, all these works still use binary relevance for training and evaluation, *i.e.* an image/caption is either relevant or not, excluding images which may be somewhat relevant.

Closest to our work, Kim *et al.* [30] explore non-binary relevance in image retrieval. They propose a log-ratio loss in order to learn a metric embedding space without requiring binary relevance between items. Their work is primarily focused on human pose, in which they use the distance between joints to rank images. They also explore within-modal image retrieval using word mover's distance, as a proxy for semantic similarity. Up to our knowledge, [30] offers the only prior work, albeit in image retrieval, to investigate both training and evaluating relevance which extends beyond both binary and instance-based relevance.

2.2. Video Retrieval

Early works in video retrieval simply extended image-retrieval approaches by temporally aggregating frames for each video [20, 52, 61, 67]. These works are attributed for defining the cross-modal video retrieval problem and standard evaluation metrics. In qualitative results, they argue models are superior if they retrieve multiple relevant videos, despite the quantitative metrics only evaluating the corresponding video.

With larger datasets becoming available [6, 31, 42, 50, 64, 66, 73], methods focused on using self-supervision [1, 59, 75], sentence disambiguation [14, 65], multi-level encodings [21, 69], mixing *"expert"* features from pre-trained models [25, 37, 41, 45] and weakly-supervised learning

from massive datasets [40, 42, 54]. All these works train and evaluate for instance-based video retrieval.

Two recent works explored using semantic similarity during training [54, 65]. Our previous work [65] uses class knowledge to cluster captions into relevance sets for triplet sampling. Patrick *et al.* [54] propose a captioning loss, where the embedding caption is re-constructed from a support set of videos. This ensures shared semantics are learned between different instances and gives large improvements when the support set does not include the corresponding video—forcing the model to generalise. However, this work is evaluated using instance-based retrieval.

This paper is the first to scrutinise current benchmarks in video retrieval, which assume instance-based correspondence. We propose semantic similarity video retrieval as an alternative task, for both evaluation and training.

3. Shortcomings of Current Video Retrieval Benchmarks

In this section, we formalise the current approaches to video retrieval, and highlight the issues present with their Instance-based Video Retrieval (IVR) assumption, which impacts the evaluation of common benchmark datasets.

Formally, given a set of videos X and a corresponding set of captions Y, current approaches define the similarity S_I between a video x_i and a caption y_j which captures this one-to-one relationship. For each video/caption there is exactly one relevant caption/video:

$$S_I(x_i, y_j) = \begin{cases} 1, & \text{if } i == j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Alternatively, if multiple captions are collected per video as in [66], then we consider the caption $y_{j,k}$ as the kth caption of the jth video. As in Eq. 1, this only considers captions of the corresponding video to be relevant.

IVR relies on the correspondence between the video and the caption captured during dataset collection. This is typically a caption provided by an annotator or transcribed from the video's narration. Importantly, the above formulation makes the assumption that no two captions of different videos are relevant-enough to impact the evaluation or training of retrieval methods. We start by qualitatively examining this assumption for current benchmarks.

Datasets In Table 1 we show the statistics of datasets that are actively being used as benchmarks for video retrieval. We order these by the size of the test set, as a larger test set is not only challenging in distinguishing between more examples, but importantly increases the chance of having other relevant items, beside the corresponding video/caption.

Most datasets [13, 31, 66, 73] have been collected from YouTube and annotated after the fact via crowd-sourcing. Notably, MPII movie [58] instead used movie scripts as captions for each of the video clips and EPIC-KITCHENS

	Text Type	# Captions	Test Set Size ↓	Scenario	Source	Eval. Metrics	Semantic Info
MSVD [13]	Caption	*86k	670	Open	YouTube	Recall@k, Avg. Rank	Multi-Lang.
MPII movie [58]	Script	64k	1,000	Movie Scripts	Movies	Recall@k, Avg. Rank	None
DiDeMo [6]	Caption	40k	1,004	Open	Flickr	Recall@k	None
MSR-VTT [66]	Caption	200k	2,990	Open	YouTube	Recall@k, Avg. Rank	Category
YouCook2 [73]	Caption	15k	3,310	Cooking	YouTube	Recall@k, Avg. Rank	None
QuerYD [50]	Caption	31k	4,717	Open	YouTube	Recall@k, Avg. Rank	Category
ActivityNet+Captions [31]	Dense Captioning	100k	4,917	Daily Living	YouTube	Recall@k	None
TVR [33]	Video Subtitle	109k	5,445	TV Shows	TV	Recall@k	TV Show
Condensed Movies [7]	Caption	34k	6,581	Movies	YouTube	Recall@k, Avg. Rank	Movie
VATEX [64]	Caption	*412k	8,920	Open	YouTube	Recall@k, Avg. Rank	Multi-Lang.
EPIC-KITCHENS [18]	Short Caption	77k	9,668	Kitchen	Egocentric	mAP,nDCG	Action Class

Table 1: Details of popular Datasets in Video Retrieval, ordered by test set size. *Number of English captions.
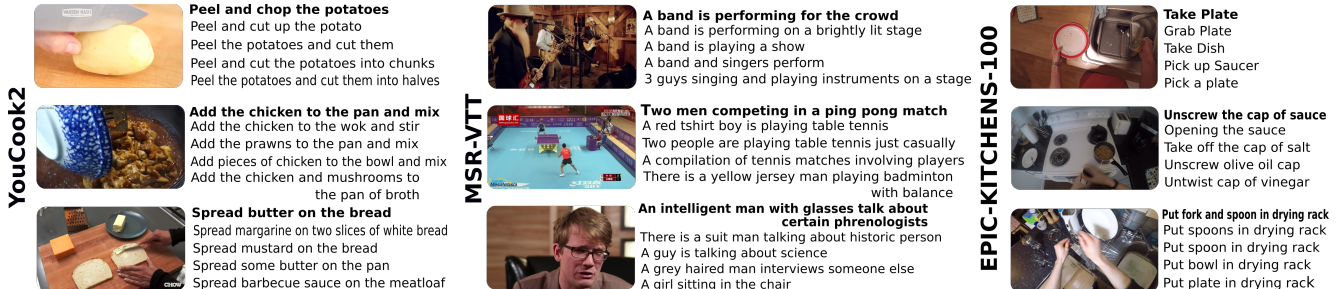

Figure 2: Video examples from the test set of three datasets showcasing the corresponding caption (bold) used as ground-truth along with highly relevant captions for other videos in the test set, considered irrelevant by IVR. In fact at times, such as the top example from MSR-VTT, a caption deemed irrelevant by the benchmarks can be a more specific description of the video.

utilised transcribed audio narrations provided by the video collectors. However, in all cases, the captions were collected with the annotator observing a single video, thus a caption's relevance to other videos could not be considered.

MSR-VTT [66], MSVD [13] and VATEX [64] include multiple captions per video, from multiple annotators, due to the datasets being collected for captioning and paraphrase evaluation. Nevertheless, during evaluation, prior works [41, 45, 69] all use a test set that considers only one caption per video. While some works [15, 25, 37, 41, 65] utilise multiple captions during training, captions are only relevant to the corresponding video and considered irrelevant to all other videos. The video pentathlon [3] recently defined a retrieval challenge across five datasets [6, 13, 31, 66, 73]. This pentathlon similarly utilises IVR.

We focus our analysis on three datasets with a large test set, MSR-VTT, YouCook2 and EPIC-KITCHENS. We consider YouCook2 and EPIC-KITCHENS as these focus on the single scenario of cooking, increasing the number of relevant captions within the dataset.

Qualitative Analysis We start by highlighting stark qualitative examples, showcasing the shortcomings of the IVR assumption, in Fig. 2. For each video, we show a key frame along with five captions from the test set. We highlight the corresponding caption according to the dataset annotations in bold—which is used as ground-truth for evaluating and ranking various methods. In each case, we show several in-distinguishable captions that are all relevant in describing the corresponding video. In fact, identifying which caption is the ground truth would be challenging for a human. However, a method that *potentially randomly* gets the bold captions higher in the retrieval list would be considered state-of-the-art, while another might be unfairly penalised. These valid captions contain synonyms, a change in the sentence structure or more/less details in describing the video.

Additionally, we find captions which are not interchangeable but are still somewhat relevant to the video. For instance, the second example of EPIC-KITCHENS includes captions of opening other bottles—*e.g.* sauce bottle vs the vinegar/oil bottle. These captions should be ranked higher than an irrelevant caption (*e.g.* "cutting a tomato").

Conclusion While the concern with IVR is clarified in this section, the task of manually annotating all relevant captions, as well as somewhat relevant captions, is unachievable due to time and cost required. Instead, we propose several proxy measures for semantic similarity between videos and captions, which require no extra annotation effort and use external corpora or knowledge bases.

4. Video Retrieval with Semantic Similarity

In this paper, we propose to move beyond Instance-based Video Retrieval (IVR) towards video retrieval that uses semantic similarity between videos and captions, for both video-to-text and text-to-video retrieval. We first define Se-

mantic Similarity Video Retrieval (SVR), then propose an evaluation protocol, as well as an approach to incorporate semantic similarity during training. Finally, in Sec. 4.4 we propose multiple approaches to estimate semantic similarity from captions without the need for manual annotations.

4.1. Definition

Given the set of videos, X, and a corresponding set of captions, Y. We define a semantic similarity function, $S_S(x_i, y_j) \rightarrow [0, 1]$, which calculates a continuous score that captures the similarity between any (video, caption) pair. Similar to IVR, $S_S(x_i, y_j) = 0$ if the caption is irrelevant to the video and 1 for maximally relevant. Different from IVR, multiple captions can have a similarity of 1 to a video, and analogously for videos. Additionally, the continuous value of S_S can model varying levels of similarity. If $S_S(x_i, y_j) > S_S(x_i, y_k)$ then y_j is a more relevant caption to the video x_i than the caption y_k. Consequently, if $S_S(x_i, y_j) = S_S(x_i, y_k)$ then both captions are considered equally relevant and retrieving them in any order should not be penalised by the evaluation metric.

4.2. Evaluation

To accommodate for cross-modal retrieval, i.e. both text-to-video and video-to-text, we use the terms "item" and "query" to refer to either a video or a caption. For a given query, all items from the opposing modality are ranked according to their distance from the query in the learnt embedding space. Benchmarks in IVR use the following evaluation metrics: Recall@K, Geometric Mean[1] and Average Rank (median or mean) of the corresponding item.

In SVR, a different evaluation metric is needed due to limitations of all current evaluation metrics used for IVR. Firstly, Average Rank only allows for a single relevant item. Whilst Recall@K can be used to evaluate queries with multiple items, a threshold on the continuous similarity is required. Additionally, choosing the value of K has to be considered carefully. If the value of K is less than the number of relevant items for a given query, the metric would not be suitable to assess a model's true performance. This is a concern for SVR where the number of relevant items will vary per query, resulting in an unbalanced contribution of different query items to the metric. Mean Average Precision (mAP) has also been used for retrieval baselines as it allows for the full ranking to be evaluated. However, mAP also requires binary relevance between query and items.

We seek an evaluation metric which is able to capture multiple relevant items and take into account relative non-binary similarity. We thus propose using normalised Discounted Cumulative Gain (nDCG) [29]. nDCG has been

used previously for information retrieval [12, 56]. It requires similarity scores between all items in the test set. We calculate Discounted Cumulative Gain (DCG) for a query q_i and the set of items Z, ranked according to their distance from q_i in the learned embedding space:

$$DCG(q_i) = \sum_{j=1}^{|\mathcal{R}_{q_i}|} \frac{2^{S_S(q_i, z_j)} - 1}{log(j + 1)} \qquad (2)$$

where $\mathcal{R}_{q_i}$ is the set of all items of the opposing modality, excluding irrelevant items, for query $q_i : \mathcal{R}_{q_i} = \{z_j | S_S(q_i, z_j) > 0, \forall z_j \in Z\}$[2]. Note that this equation would give the same value when items of the same similarity S_S are retrieved in any order. It also captures differing levels of semantic similarity.

nDCG can then be calculated by normalising the DCG score such that it lies in the range $[0, 1]$: $nDCG(q_i) = \frac{DCG(q_i)}{IDCG(q_i)}$ where $IDCG(q_i)$ is calculated from DCG and Z ordered by relevance to the query q_i.

For overall evaluation, we consider both video-to-text and text-to-video retrieval and evaluate a model's nDCG as:

$$nDCG(X, Y) = \frac{1}{2|X|} \sum_{x_i \in X} nDCG(x_i) + \frac{1}{2|Y|} \sum_{y_i \in Y} nDCG(y_i)$$
$$(3)$$

Note that Eq. 3 allows for a different number of videos and captions in the test set.

4.3. Training

In addition to utilising semantic similarity for evaluation, it can also be incorporated during training. A contrastive objective can be defined to learn a multi-modal embedding space, e.g. the triplet loss:

$$L_t(x_i, y_j, y_k) = \max\left(m + D(f(x_i), g(y_j)) - D(f(x_i), g(y_k)), 0\right)$$
$$(4)$$

where $D(\cdot, \cdot)$ is a distance function, $f(\cdot)$ and $g(\cdot)$ are embedding functions for video and text respectively, and m is a constant margin. In IVR, the triplets x_i, y_j and y_k are sampled such that $S_I(x_i, y_j) = 1$ and $S_I(x_i, y_k) = 0$ (see Eq. 1). In SVR, we use triplets such that $S_S(x_i, y_j) \geq T$ and $S_S(x_i, y_k) < T$ where T is a chosen threshold.

Alternative Losses Other alternatives to the triplet loss can be utilised, such as the approximate nDCG loss from [55], log-ratio loss from [30], or losses approximating mAP [10, 11, 55, 57]. It is worth noting that some of these works combine the proposed loss with the instance-based triplet loss for best performance [10, 30]. Additionally, approximating mAP requires thresholding as mAP expects binary relevance. Note that all these works, apart from [30], attempt instance-based image retrieval. Experimentally, we found the log-ratio loss to produce inferior results to thresholding the triplet loss. Adapting these losses to the SVR task is an exciting area for exploration in future work.

---

[1]Geometric Mean averages Recall@K over a range, typically $\{1, 5, 10\}$, each giving the percentage of queries for which the corresponding item was found within the top K results.

[2]Note that nDCG does not penalise the case when a large number of low-relevant items are present. This can be alleviated by thresholding S.

4.4. Proxy Measures for Semantic Similarity

Collecting semantic similarity from human annotations, for all but the smallest datasets, is costly, time consuming[3] and potentially noisy. Previous work in image retrieval [26] demonstrated that semantic similarities of captions can be successfully utilised. We use the knowledge that each video in the dataset was captured with a corresponding caption, which offers a suitable description of the video, and thus use the semantic similarity between captions instead, *i.e.* we define $S_S(x_i, y_j)$ as

$$S_S(x_i, y_j) = \begin{cases} 1 & i == j \\ S'(y_i, y_j) & otherwise \end{cases} \quad (5)$$

where S' is a semantic proxy function relating two captions.

We define four semantic similarity measures which we use to compute $S'(y_i, y_j)$—based on bag of words, part-of-speech knowledge, synset similarity and the METEOR metric [8]. We choose these proxy measures such that they should scale with the size of the dataset, not requiring any extra annotation effort, but acknowledge that some datasets may be better suited by one proxy over others. We investigate this qualitatively and quantitatively in Sec. 5.1.

Bag-of-Words Semantic Similarity Naively, one could consider the semantic similarity between captions as the overlap of words between them. Accordingly, we define the Bag-of-Words (BoW) similarity as the Intersection-over-Union (IoU) between sets of words in each caption:

$$S'_{BoW}(y_i, y_j) = \frac{|w_i \cap w_j|}{|w_i \cup w_j|} \quad (6)$$

where w_i and w_j represent the sets of words, excluding stop words, corresponding to captions y_i and y_j.

This proxy is easy to calculate, however, as direct word matching is used with no word context. This raises two issues: firstly, synonyms for words are considered as irrelevant as antonyms, *i.e. "put"* and *"place"*. Secondly, words are treated equally—regardless of their part-of-speech, role in the caption, or how common they are. Word commonality is partially resolved by removing stop words[4]. We address the other concerns next.

Part-of-Speech Semantic Similarity Verbs and nouns, as well as adjectives and adverbs, describe different aspects of the video and as such words can be matched within their part-of-speech. Matching words irrespective of part-of-speech can result in incorrect semantic similarities. For example, the captions "watch a play" and "play a board game". Alternatively, adverbs can be useful to determine how-to similarities between captions [22]. By augmenting the part-of-speech, we can ensure that the actions and objects between two videos are similar.

To calculate the Part-of-Speech (PoS) word matching, captions are parsed, and we calculate the IoU between the sets of words for each of the parts-of-speech and average over all parts-of-speech considered.

$$S'_{PoS}(y_i, y_j) = \sum_{p \in P} \alpha^p \frac{|w_i^p \cap w_j^p|}{|w_i^p \cup w_j^p|} \quad (7)$$

where p is a part-of-speech from the set P, w_i^p is the set of words from caption y_i which have a part-of-speech p, and α^p is the weight assigned to p such that $\sum_{p \in P} \alpha^p = 1$.

Synset-Aware Semantic Similarity So far, the proxies above do not account for synonyms, *e.g. "put"* and *"place"*, *"hob"* and *"cooker"*. We extend the part-of-speech similarity detailed above using semantic relationship information from synsets, i.e. grouped synonyms, from WordNet [44] or other semantic knowledge bases. We modify the part-of-speech proxy,

$$S'_{SYN}(y_i, y_j) = \sum_{p \in P} \alpha^p \frac{|C_i^p \cap C_j^p|}{|C_i^p \cup C_j^p|} \quad (8)$$

where C_i^p is the set of synsets within the part-of-speech p for caption y_i. Note that $|C_i^p| \leq |w_i^p|$ as multiple words are assigned to the same synset due to the similar meanings.

METEOR Similarity The first three similarity functions break the sentence into its individual words, with/without parsing knowledge. Instead, captioning works have proposed metrics that preserve the structure of the sentence, comparing two captions accordingly. Multiple metrics have been proposed (e.g. BLEU [53], ROUGE [36] or CIDEr [62]) including METEOR [8]. Originally used for machine translation and later image captioning, Gordo and Larlus [26] proposed METEOR as one of their proxy measures for relating images via their captions.

METEOR calculates similarity both via matching, using synsets to take into account synonyms, and via sentence structure, by ensuring that matched words appear in a similar order. The proxy is then defined as: $S'_{MET}(y_i, y_j) = M(y_i, y_j)$, where $M(\cdot, \cdot)$ is the METEOR scoring function.

Other proxies Other similarity measures, including the use of word/sentence embedding models such as BERT [19], do not provide useful similarity scores on video retrieval datasets. This is further discussed in supplementary.

5. Semantic Similarity Analysis

We evaluate baseline methods on the three datasets, with the aim of answering the following questions: (i) How do the different proxy measures compare to each other on the three datasets? (ii) What is the impact of the noted shortcomings of IVR on methods' performance? (iii) How do current methods perform when using SVR evaluation for the four proposed proxy measures? (iv) How does training the models for SVR affect the results?

---

[3]Annotators would have to observe a video with two captions and indicate their relative relevance. For n videos and m captions this is $O(nm^2)$.

[4]We find using tf-idf to remove/re-weight words comparable to removing stop words in the analysed datasets.

YouCook2

	BoW	PoS	SYN	MET
heat some oil in a deep pan and add chopped onions and fry till they turn brown	1.0	1.0	1.0	1.0
add chopped onions to a pan of oil	0.42	0.7	0.75	0.32
heat oil and add chopped onion	0.42	0.63	0.71	0.81
heat oil in a pan	0.25	0.43	0.46	0.72
add chopped carrot and celery	0.14	0.2	0.25	0.41
soak the mushrooms in nearly boiling water	0.0	0.0	0.0	0.031

	BoW	PoS	SYN	MET
mix the ingredients in the pan together	1.0	1.0	1.0	1.0
stir all of the ingredients in the pan	0.5	0.5	1.0	0.7
stir the food in the pan	0.2	0.17	1.0	0.54
add the chicken to the pan and mix	0.4	0.42	0.42	0.44
fry the chicken in the pan	0.2	0.17	0.17	0.61
crush some garlic	0.0	0.0	0.0	0.0

MSR-VTT

	BoW	PoS	SYN	MET
a girl and the judges talking on the voice	1.0	1.0	1.0	1.0
a female judge is speaking to a young girl on a televised music competition	0.8	0.83	1.0	0.64
a contestant is judged on a tv show	0.4	0.33	0.17	0.38
a boy talking to competition judges	0.5	0.75	0.62	0.6
there is a woman is talking with two guys	0.25	0.35	0.6	0.33
a dog is playing	0.0	0.0	0.0	0.14

	BoW	PoS	SYN	MET
a group of men wrestling	1.0	1.0	1.0	1.0
a wrestling match is going on	0.8	0.88	0.5	0.62
2 guys wrestling each other	0.6	0.75	0.5	0.67
a song plays while a people compete at a wrestling meet	0.5	0.7	0.83	0.45
a man is holding another man from behind	0.67	0.8	0.62	0.43
a boy and a girl are hugging a man dressed as santa	0.0	0.0	0.0	0.12

EPIC-KITCHENS-100

	BoW	PoS	SYN	MET
pour olive oil in pan	1.0	1.0	1.0	1.0
pour oil in pan	0.5	0.67	1.0	0.6
pour olive oil in plate	0.67	0.67	0.67	0.79
put in pan	0.33	0.25	0.25	0.39
close olive oil	0.33	0.25	0.25	0.39
mix thyme and olive oil in plate	0.33	0.12	0.12	0.57

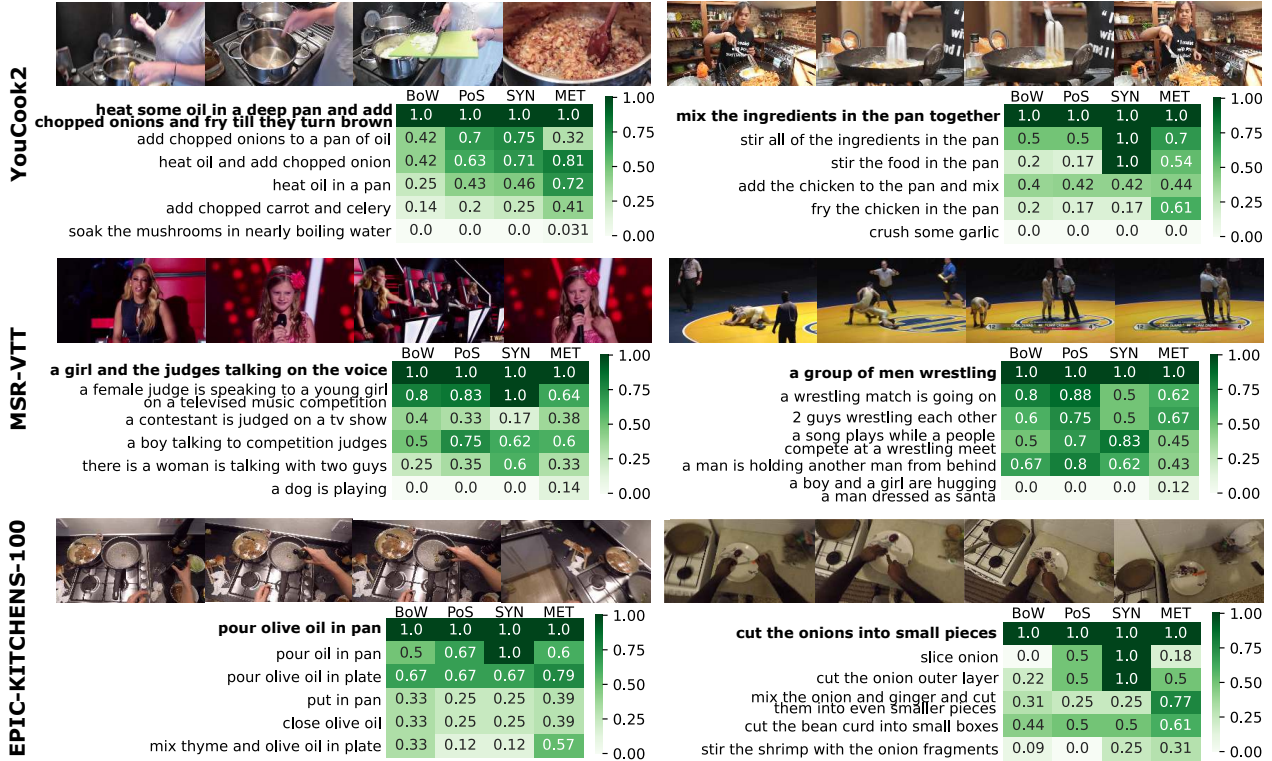|	BoW	PoS	SYN	MET
cut the onions into small pieces	1.0	1.0	1.0	1.0
slice onion	0.0	0.5	1.0	0.18
cut the onion outer layer	0.22	0.5	1.0	0.5
mix the onion and ginger and cut them into even smaller pieces	0.31	0.25	0.25	0.77
cut the bean curd into small boxes	0.44	0.5	0.5	0.61
stir the shrimp with the onion fragments	0.09	0.0	0.25	0.31

Figure 3: Examples of the proposed semantic similarity proxies (Sec 4.4). Captions are shown alongside the score $S'_S(y_i, y_j)$ when compared to the corresponding caption (bold). While scores differ, methods agree on highly-(ir)relevant captions.

Next, we present information on the datasets and baseline methods along with their implementation details.

Datasets We continue exploring the three public datasets from Sec. 3. These are: frequently-used MSR-VTT [66] and YouCook2 [73], as well as recently released EPIC-KITCHENS-100 [18]. The latter also has the benefit in that it offers semantic annotations as we show next.

Baselines We train a simple embedding baseline with a multi-layer perceptron for each modality which we name as Multi-Modal Embedding or **MME**. We additionally consider three publicly available baselines for the aforementioned datasets. We use the benchmark implementations provided by the video pentathlon challenge [3] for MSR-VTT and YouCook2: **MoEE [41]:** Multiple video features are extracted from 'video experts' and an embedding is learned for each. The final embedding is learned as a weighted sum, determined by the caption. **CE [37]:** Video expert features are passed through a paired collaborative gating mechanism before the embedding and resulting weighted sum. For EPIC-KITCHENS, we use the baseline method **JPoSE [65]:** this trains separate embedding spaces for each part-of-speech in the caption before being combined into a retrieval embedding space. Implementation details match the publicly available code per baseline as trained for IVR.

Parsing and Semantic Knowledge We parse the captions using Spacy's large web model [24]. We limit these to verbs and nouns, setting $\alpha^p = 0.5$ for each in all experiments. When computing the Synset-Aware Similarity, we use the synsets released as part of [18] for both EPIC-KITCHENS and YouCook2, as both share the domain of cooking. We found that the synset information transfers well across both datasets. Synset knowledge for MSR-VTT is found using WordNet [44] and the Lesk algorithm [34]. MSR-VTT includes multiple captions per video, therefore, for robust word sets, we only include words which are present in 25% or more of all the captions for a given video (excluding stop words). For METEOR, we use the NLTK implementation [38]. Additionally, to calculate S_{MET} for MSR-VTT, we use many-to-many matching with a non-Mercer match kernel [39].

5.1. Proxy Measure Comparisons

We first clarify differences between the semantic similarity proxies with qualitative examples. Fig. 3 shows examples from YouCook2, EPIC-KITCHENS and MSR-VTT.

BoW is the tightest proxy to IVR, only considering captions as equally relevant when the set of words match exactly. The Synset proxy is the only one to consider the captions "stir food in the pan" and "mix the ingredients in the pan together" equivalent. This is because it separately focuses on the verb and noun (similar to PoS) and is able to
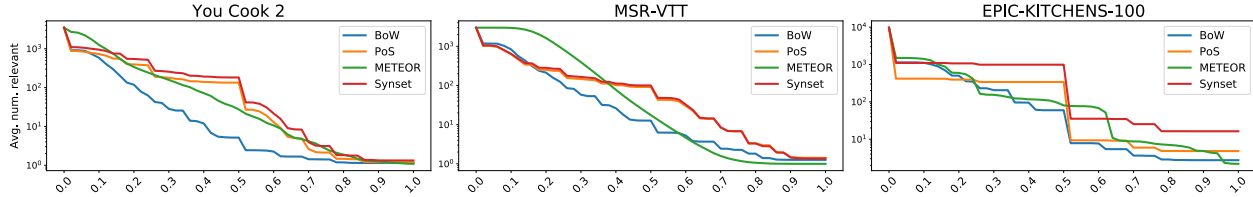
Figure 4: Average number of relevant captions for a video with a given threshold over each dataset and proxy measure.
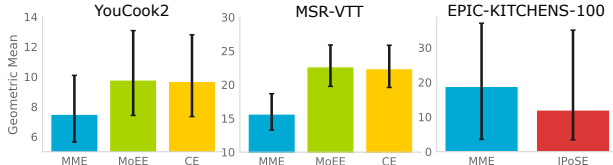

Figure 5: The min. and max. performance of baseline methods on the instance-based metric geometric mean when considering captions with $S'_{SYN}(y_i, y_j) > 0.8$ equivalent.

relate words such as *"stir"* and *"mix"*. While METEOR also considers synonyms, it aims for a good alignment in word order, therefore it gives all captions containing *"in the pan"* a high score, even when the verb differs. This also explains the low score given to "add chopped onions to a pan of oil" compared to PoS and SYN even though the caption contains many of the same concepts.

For MSR-VTT, we show examples that demonstrate limitations of semantic proxies. All proxies rank the caption *"a boy talking..."* higher than *"a contestant is judged..."*. Similarly, the relevance of the caption *"a song plays"* requires access to the audio associated with the video and cannot be predicted from the associated caption.

Having established an understanding of the proxies, we now quantitatively assess them. In Fig. 4, we calculate the similarity between a video and all captions in the dataset using Eq. 10. We then vary the threshold, T, for each proxy and compute the number of captions where $S(x_i, y_j) \geq T$. We plot the average number of 'relevant' captions as the threshold increases. Note the y-axis is a log scale. In all cases, we note that even at high thresholds the average number of relevant captions is higher than 1. As expected, the synset proxy, includes as many or more relevant captions than PoS, due to it considering synonyms as equivalent. This is most evident for EPIC-KITCHENS.

5.2. Shortcomings of IVR Evaluation

In Sec. 3, we analysed the shortcomings of the current approach to video retrieval that only considers a single relevant caption—the corresponding one. In this section, we use the semantic proxies to quantify the impact of IVR on the evaluation of video retrieval methods.

We consider the Geometric Mean metric, used as the prime metric in the pentathlon [3]. For each method, we showcase an upper/lower limit (as an error bar). To cal-

culate this we consider the retrieved captions and locate the highest-retrieved caption that is above a tight threshold $S'_{SYN}(y_i, y_j) > 0.8$, per video (see Fig. 3 for examples). We re-calculate the metrics, and show this as an upper limit for the method's performance. Similarly, we locate the lowest-retrieved caption above the same threshold. This provides the lower limit. The figure shows the error in the evaluation metric, for each baseline on all datasets.

From Fig. 5 we demonstrate a significant change in Geometric Mean when using the Synset-Aware proxy ($\sim$30 geometric mean for EPIC-KITCHENS, $\sim$6.0 for MSR-VTT and $\sim$5.0 for YouCook2). The gap between the reported performance and the upper-bound indicates that these baselines are retrieving some highly similar captions as more relevant than the ground-truth. Instance-based evaluation metrics do not account for this. Without considering this analysis on all relevant captions, we believe it is not possible to robustly rank methods on these benchmarks.

5.3. Using Semantic Proxies for Evaluation

We now evaluate SVR using nDCG (Eq. 3) with our proposed semantic similarity proxies. Without re-training, we evaluate nDCG on the test set, where the semantic similarities are defined using one of the four proxies in Sec. 4.4. We present the results in Fig. 7 on the three datasets.

Baselines significantly outperform Random as well as the simple MME baseline on instance-based Geometric Mean. However, when semantic similarity proxies are considered, this does not hold. For almost all cases, MoEE, CE and JPoSE are comparable to MME. MME even outperforms more complex approaches (e.g. on YouCook2). This is critical to demonstrate, as proposed methods can produce competitive results on the problematic IVR setup, but may not have the same advantage in SVR.

In Fig. 7 we can also see that the METEOR proxy leads to high nDCG values even for the Random baseline on MSR-VTT and YouCook2. This is due to high inter-caption similarities on average. Differently, JPoSE outperforms MME and Random on EPIC-KITCHENS for the METEOR proxy. This suggests the hierarchy of embeddings in JPoSE improves the sentence structure matches.

While the various proxies differ in the scores they assign to captions, all four are suitable to showcase that tested baselines do not improve over MME. This demonstrates that, regardless of the semantic proxy, it is important to con-
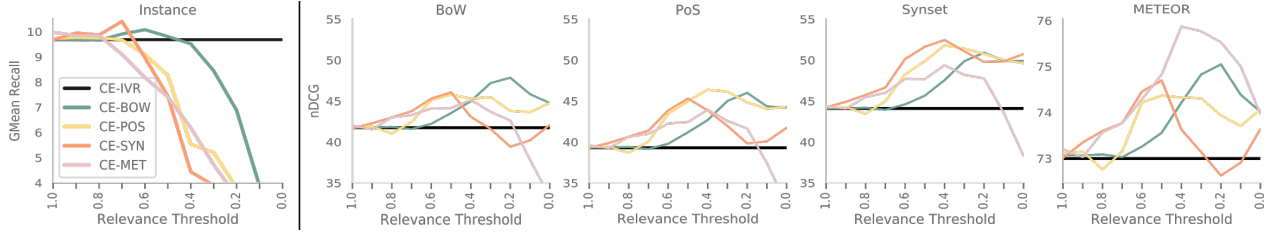
Figure 6: Training CE with semantic knowledge compared to instance-only on (left) IVR using Geometric Mean and (right) the four proposed semantic proxies using nDCG. Using semantic proxy in training improves performance in every case.
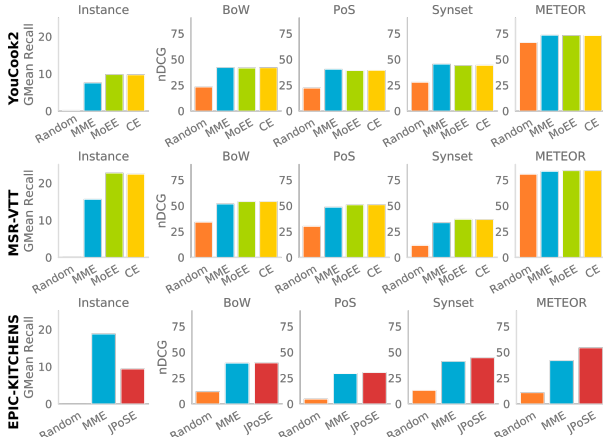

Figure 7: Evaluating the baseline methods on the proxy measures for semantic similarity (Table in supplementary).

sider semantic similarity when assessing a method's performance, rather than using IVR.

Choice of Semantic Proxy We consider all four proposed proxies to be valuable similarity metrics. One proxy can be chosen over another for certain objectives/applications. For example, to retrieve videos of the same recipe, BoW is useful as only videos containing the same step and ingredients are considered highly relevant. Conversely, PoS and SYN are useful when focusing on actions as they increase the importance of verbs. SYN is also particularly useful for free form captions, where synonyms are plentiful. Multiple proxies can be considered as multiple evaluation metrics for increased robustness.

5.4. Training with Semantic Similarity

So far, the models have been trained solely using current IVR losses. We now train with semantic knowledge using the method from Sec. 4.3. We limit these experiments to YouCook2, and the Collaborative Experts [37] (CE) baseline due to the number of models required for training for each proxy measure and threshold T. We use the following labels to refer to models trained with the four proxy measures: *CE-BoW*, *CE-PoS*, *CE-SYN* and *CE-MET* respectively. The original model trained using IVR, is designated as *CE-IVR*. We vary the threshold $T = \{0.1, 0.2, ..., 1\}$, showing the results in Fig. 6 for both IVR (left) and SVR

(right). All plots compare to the *CE-IVR* (black line).

Fig. 6 (left) demonstrates that for all proxies, providing semantic information during training can increase the performance of IVR, however this does drop off as less similar items are treated as relevant. As anticipated, the drop-off threshold varies per semantic proxy.

Fig. 6 (right) shows that as T decreases, and more captions are considered relevant in training, significant improvement in nDCG can be observed compared to *CE-IVR*. Note that the nDCG value cannot be compared across plots, as these use different semantic proxies in the evaluation. While the highest performance is reported when considering the same semantic proxy in both training and evaluation, training with any proxy improves results, although they peak at different thresholds. From inspection, *CE-SYN*, *CE-MET* and *CE-PoS* peak in performance around $T = 0.4$ whereas *CE-BoW* has a peak at $T = 0.2$. When training with these specific thresholds, the models are able to best learn a semantic embedding space, which we find is agnostic of the proxy used in evaluation.

6. Conclusion

This paper highlights a critical issue in video retrieval benchmarks, which only consider instance-based (IVR) similarity between videos and captions. We have shown experimentally and through examples the failings of the assumption used for IVR. Instead, we propose the task of Semantic Similarity Video Retrieval (SVR), which allows multiple captions to be relevant to a video and vice-versa, and defines non-binary similarity between items.

To avoid the infeasible burden of annotating datasets for the SVR task, we propose four proxies for semantic similarity which require no additional annotation effort and scale with dataset size. We evaluated the proxies on three datasets, using proposed evaluation and training protocols. We have shown that incorporating semantic knowledge during training can greatly benefit model performance. We provide a public benchmark for evaluating retrieval models on the SVR task for the three datasets used in this paper at: https://github.com/mwray/Semantic-Video-Retrieval.

References

[1] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *CoRR*, abs/2006.16228, 2020. 2

[2] Samuel Albanie, Yang Liu, Arsha Nagrani, Antoine Miech, Ernesto Coto, Ivan Laptev, Rahul Sukthankar, Bernard Ghanem, Andrew Zisserman, Valentin Gabeur, et al. The end-of-end-to-end: A video understanding pentathlon challenge benchmark. https://www.robots.ox.ac.uk/~vgg/challenges/video-pentathlon/. Accessed: 16th Nov. 2020. 1

[3] Samuel Albanie, Yang Liu, Arsha Nagrani, Antoine Miech, Ernesto Coto, Ivan Laptev, Rahul Sukthankar, Bernard Ghanem, Andrew Zisserman, Valentin Gabeur, et al. The end-of-end-to-end: A video understanding pentathlon challenge (2020). *CoRR*, abs/2008.00744, 2020. 3, 6, 7

[4] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *ECCV*, 2018. 1

[5] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 1

[6] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2, 3

[7] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 2020. 3

[8] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACLW*, 2005. 2, 5

[9] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *WACV*, 2019. 2

[10] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-AP: Smoothing the path towards large-scale image retrieval. In *ECCV*, 2020. 4

[11] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *CVPR*, 2019. 4

[12] Olivier Chapelle and Mingrui Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 2010. 4

[13] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 2, 3

[14] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. Cross-modal image-text retrieval with semantic consistency. In *ACM-MM*, 2019. 2

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

[16] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 1

[17] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 2

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020. 1, 3, 6

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 5, 11

[20] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Trans. Multimed.*, 2018. 2

[21] Jianfeng Dong, Xirong Li, Chaoxi Xu, Gang Yang, and Xun Wang. Hybrid space learning for language-based video retrieval. *TPAMI*, 2020. 2

[22] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action Modifiers: Learning from Adverbs in Instructional Videos. In *CVPR*, 2020. 5

[23] Peter GB Enser, Christine J Sandom, Jonathon S Hare, and Paul H Lewis. Facing the reality of semantic image retrieval. *Journal of documentation*, 2007. 2

[24] Explosion. Spacy. https://spacy.io/. Accessed: 6th Nov. 2020. 6

[25] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2, 3

[26] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, 2017. 2, 5

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1

[28] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. 1

[29] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 2002. 4

[30] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. Deep metric learning beyond binary supervision. In *CVPR*, 2019. 2, 4

[31] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 2, 3

[32] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015. 11

[33] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVR: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 3

[34] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC*, 1986. 6

[35] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *CVPR*, 2019. 1

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACLW*, 2004. 5

[37] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 2, 3, 6, 8

[38] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *CoRR*, abs/0205028, 2002. 6

[39] Siwei Lyu. Mercer kernels for object recognition with local features. In *CVPR*, 2005. 6

[40] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2

[41] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *CoRR*, abs/1804.02516, 2018. 2, 3, 6

[42] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1, 2

[43] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 2, 11

[44] George A Miller. Wordnet: a lexical database for english. *Commun. ACM*, 1995. 2, 5, 6

[45] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*, 2018. 2, 3

[46] Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas, and Dima Damen. Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In *ICCV*, 2017. 1

[47] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, 2020. 1

[48] T Ojala, M Rautiainen, E Matinmikko, and M Aittola. Semantic image retrieval with hsv correlograms. In *SCIA*, 2001. 2

[49] Aude Oliva, Antonio B Torralba, Anne Guérin-Dugué, and Jeanny Hérault. Global semantic classification of scenes using power spectrum templates. In *Challenge of image retrieval*, 1999. 1

[50] Andreea-Maria Oncescu, Jõao F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality textual and audio narrations. In *ICASSP*, 2021. 2, 3

[51] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *BMVC*, 2020. 1

[52] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *ECCV*. Springer, 2016. 2

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[54] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2020. 2

[55] Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. *Information retrieval*, 2010. 4

[56] Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. In *ACM SIGIR*, 2010. 4

[57] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019. 4

[58] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 1, 2, 3

[59] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. Avlnet: Learning audio-visual language representations from instructional videos. *CoRR*, abs/2006.09199, 2020. 2

[60] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 11

[61] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *CoRR*, abs/1609.08124, 2016. 2

[62] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 5

[63] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015. 1

[64] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 2, 3

[65] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 2019. 2, 3, 6

[66] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1, 2, 3, 6

[67] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015. 2

[68] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015. 1

[69] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 2, 3

[70] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *CoRR*, abs/1611.04021, 2016. 1

[71] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *CVPR*, 2016. 1

[72] Luowei Zhou. https://github.com/LuoweiZhou/YouCook2-Leaderboard. Accessed: 16th Nov. 2020. 1

[73] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. *CoRR*, abs/1703.09788, 2017. 1, 2, 3, 6

[74] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *IJCV*, 2017. 1

[75] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 2

Supplementary Material

Here we provide a perceptual study that correlates the proxy measures to human annotators in Section A. Next, we provide information of the correlation between the semantic proxies in Section B, then details on three other proxy measures showcasing their unsuitability to the three datasets in Section C. Finally, we show a tabular version of Figure 7 from the main paper for numerical comparison in future works in Section D.

A. Proxy Measures Human Agreement Study

One might wonder how do the proposed proxies in Section 4.4 correlate to human annotators. To answer this question, we conduct a small-scale human study.

Requesting a human to assign a score relating a video and a caption is challenging and potentially subjective, however, ranking a small number of captions for their relevance to a given video can be achieved. We randomly select 100 videos from both the YouCook2 and MSR-VTT datasets (we focus on these two datasets as they include the most varied captions). For each proposed proxy, we rank the corresponding captions by their similarity to a given video, then select the most/least relevant captions as well as the captions at the 1st, 2nd and 3rd quartiles. This gives us 5 captions that are semantically distinct for the video.

We then asked 3 annotators (out of 6 total annotators) to order these 5 captions by their similarity to the given video. We remove annotation noise by only considering consistently ordered pairs of captions—that is when all 3 annotators agree that caption A is more relevant than B. We then report the percentage of correctly ordered pairs by the proxy, out of all consistently annotated pairs, as the 'Human-to-Proxy' agreement.

Table 2 shows the results of this human study. We note the % of consistent pairs of captions in each case. Results demonstrate that the four proxies correlate well with human rankings, with SYN and BoW giving the best Human-to-Proxy agreement on YouCook2 and MSR-VTT respec-

	BoW	PoS	SYN	MET
% Consistent Pairs YouCook2	86.5	78.0	76.3	77.3
% Consistent Pairs MSR-VTT	73.1	78.8	75.6	69.2
Human Agreement YouCook2	91.2	88.8	92.1	85.6
Human Agreement MSR-VTT	93.7	84.8	89.7	87.5

Table 2: Human Study reporting % of caption pairs with agreement between human and proxy on YouCook2 and MSR-VTT. Note: chance is 50%.

tively. MET has lower agreements than the other proxy measures due to it penalizing different word orders as discussed in Sec. 5.1 of the main paper.

B. Correlation Between Semantic Proxies

To determine how similar the four proposed semantic proxies are, we calculate the Pearson correlation coefficient between pairs of semantic proxies for each video in YouCook2, MSR-VTT and EPIC-KITCHENS.

Figure 8 shows this correlation averaged over the videos within a dataset. All proposed semantic proxies have positive correlations, ranging between moderate (0.5-0.7) and high (> 0.7) correlations. We find the agreement between semantic proxies to be stronger at the lower end of the rank with the different methods consistently agreeing on which captions are irrelevant. At the higher end of the rank there tends to be some disagreements between proxies, with SYN and METEOR having the lowest correlation while BoW and PoS having the highest correlation. Importantly, the trend is consistent across the three datasets.

C. Proxies from Learnt Models
C.1. Definition

We compare our proposed proxies (Sec 4 in the main paper) to three other proxies which use learnt features from visual or textual models. Each proxy is defined as the cosine similarity between two vectors:

$$S'(y_i, y_j) = \frac{a(y_i) \cdot a(y_j)}{||a(y_i)|| \times ||a(y_j)||} \qquad (9)$$

where $a(\cdot)$ is a trained model.

Textual Similarity We use two language models common in the literature to get representations: Word2Vec [43] and BERT [19]. For Word2Vec, the word vectors are averaged for a sentence-level representation[5]. When using BERT, we extracted a sentence-level representation using the DistilBERT model from [60].

Visual Similarity For the visual embedding proxy, we use the video features extracted from the pre-trained model. This changes Eq. 5 in the main paper to the following:

$$S_S(x_i, y_j) = \begin{cases} 1 & i == j \\ S''(x_i, x_j) & otherwise \end{cases} \qquad (10)$$

[5]We also tried using the Word Mover's Distance [32] but achieved slightly worse results.
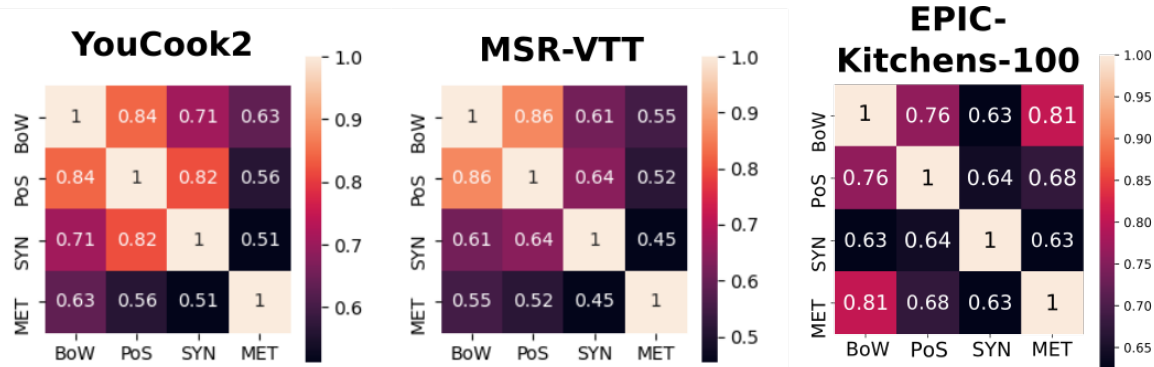
Figure 8: The average Pearson's correlation coefficient between pairs of proposed semantic proxies for YouCook2, MSR-VTT and EPIC-KITCHENS.
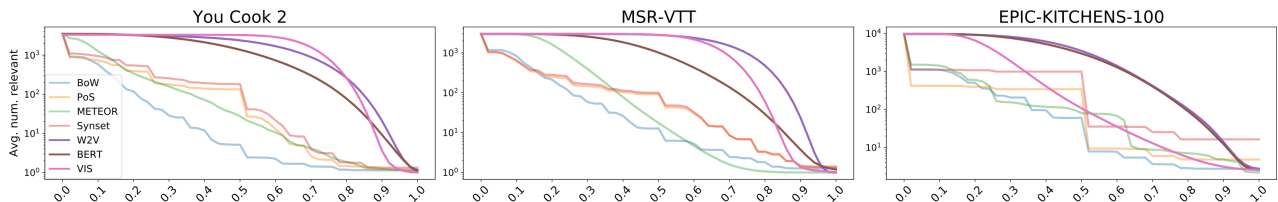

Figure 9: Average number of relevant captions for a video with a given threshold over each dataset and proxy measure including the Word2Vec (W2V), BERT and Visual (VIS).

Note that a video and a caption are related here purely on the similarity between the video features, making the assumption that the visual contents of video x_j offer a sufficient description of the caption y_j, and that the pre-trained video features offer sufficient discrimination between the videos.

C.2. Proxy Measure Comparisons

We show an extended version of Figure 4 from the main paper, adding the three proxy measures from learnt models in Figure 9. We compare these for the three datasets YouCook2, MSR-VTT and EPIC-KITCHENS.

We find the average number of relevant captions per video from the three learned proxies is much higher than the proposed proxies across almost all thresholds. With lots of captions being considered relevant, this has the effect of inflating nDCG scores.

When analysing the visual proxy, we find that the similarity is not semantic in nature. The visual proxy has high similarities between segments from the same video, further highlighting its unsuitability. Accordingly, using visual similarity from pre-trained models is not suitable as a proxy for semantic similarity.

The BERT and Word2Vec proxies similarly do not produce reasonable proxies of semantic similarities for these three datasets. From Figure 9, both methods produce significantly more relevant captions than proposed metrics. When analysing the results, we note that BERT and Word2Vec relate captions via their context, because of their training which relates words by the co-occurrence rather than their

semantic relevance. For example, 'open' and 'close' are often used in the same context of objects, but represent opposite actions. Both Word2Vec and BERT would give much higher similarity to these two, despite being antonyms.

D. Table of Figure 7

Table 3 shows the performance of the different baseline models on all three datasets and proxy measures. See Section 5.3 in the main paper for the discussion of results.

	Proxy	Instance	BoW	PoS	Syn	Met
	Metric	GMR		nDCG		
YouCook2	Random	0.1	23.1	22.1	27.7	66.2
	MEE	7.5	42.1	40.3	45.3	73.3
	MoEE	9.8	41.5	39.1	44.0	73.0
	CE	9.7	41.8	39.3	44.1	73.0
MSR-VTT	Random	0.2	34.0	30.0	11.6	80.4
	MEE	15.7	51.6	48.5	33.5	83.3
	MoEE	22.7	53.9	50.8	36.8	83.9
	CE	22.4	54.0	50.9	36.7	84.0
EPIC	Random	0.0	11.7	4.5	10.7	13.0
	MEE	18.8	39.3	29.2	41.8	41.0
	JPoSE	9.4	39.5	30.2	49.0	44.5

Table 3: Tabular version of Figure 7 from the main paper. Results of evaluating the baseline methods on the different proxy measures for semantic similarity. (GMR=Geometric Mean Recall)