

Relevance-based Margin for Contrastively-trained Video Retrieval Models

Alex Falcon
falcon.alex@spes.uniud.it
Fondazione Bruno Kessler
Trento, Italy
University of Udine
Udine, Italy

Swathikiran Sudhakaran
swathikirans@gmail.com
Samsung AI Center Cambridge
Cambridge, United Kingdom

Giuseppe Serra
giuseppe.serra@uniud.it
University of Udine
Udine, Italy

Sergio Escalera
sergio@maia.ub.es
University of Barcelona and
Computer Vision Center
Barcelona, Spain

Oswald Lanz
lanz@inf.unibz.it
Free University of Bozen-Bolzano
Bolzano, Italy

ABSTRACT

Video retrieval using natural language queries has attracted increasing interest due to its relevance in real-world applications, from intelligent access in private media galleries to web-scale video search. Learning the cross-similarity of video and text in a joint embedding space is the dominant approach. To do so, a contrastive loss is usually employed because it organizes the embedding space by putting similar items close and dissimilar items far. This framework leads to competitive recall rates, as they solely focus on the rank of the groundtruth items. Yet, assessing the quality of the ranking list is of utmost importance when considering intelligent retrieval systems, since multiple items may share similar semantics, hence a high relevance. Moreover, the aforementioned framework uses a fixed margin to separate similar and dissimilar items, treating all non-groundtruth items as equally irrelevant. In this paper we propose to use a variable margin: we argue that varying the margin used during training based on how much relevant an item is to a given query, i.e. a relevance-based margin, easily improves the quality of the ranking lists measured through nDCG and mAP. We demonstrate the advantages of our technique using different models on EPIC-Kitchens-100 and YouCook2. We show that even if we carefully tuned the fixed margin, our technique (which does not have the margin as a hyper-parameter) would still achieve better performance. Finally, extensive ablation studies and qualitative analysis support the robustness of our approach. Code will be released at <https://github.com/aranciokov/RelevanceMargin-ICMR22>.

CCS CONCEPTS

• Information systems → Information retrieval; • Computing methodologies → Artificial intelligence.

KEYWORDS

deep learning, cross-modal retrieval, video retrieval, relevance

ICMR '22, June 27–30, 2022, Newark, NJ, USA

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22)*, June 27–30, 2022, Newark, NJ, USA, <https://doi.org/10.1145/3512527.3531395>.

ACM Reference Format:

Alex Falcon, Swathikiran Sudhakaran, Giuseppe Serra, Sergio Escalera, and Oswald Lanz. 2022. Relevance-based Margin for Contrastively-trained Video Retrieval Models. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22)*, June 27–30, 2022, Newark, NJ, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3512527.3531395>

1 INTRODUCTION

With the rapid growth of digital media shared on the web it becomes increasingly important for real-world applications to offer flexible, user friendly modalities to access media content at scale. Google video search for example, translates a natural language query into a ranked list of content-related videos from the web. Natural free form, unrestricted language enables a user to express the fine-grained details in an articulated query, and each user can do so with its own expressivity. Thus, a same retrieval response can be triggered with syntactically different but semantically coherent queries. This poses significant challenges to the current state of the art in cross-modal retrieval research.

Recent approaches which deal with cross-modal video retrieval aim at learning a joint embedding space [Chen et al. 2020b; Croitoru et al. 2021; Dong et al. 2021a; Wang et al. 2021] by means of contrastive losses [Hadsell et al. 2006; Miech et al. 2020; Oord et al. 2018; Schroff et al. 2015], which put the associations available in the dataset (e.g. a video and its natural language description) as close as possible while enforcing a separation margin to all the other items (see lower left of Fig. 1). During inference, the ranking list for a given query is produced by computing similarity scores with respect to all the items by means of, e.g. the dot product or the cosine similarity. By measuring the performance of the video retrieval system with rank-unaware metrics, such as recall rates, increasingly better solutions to this problem were proposed. In fact, contrastive losses synergize well with recall rates, given how they maximize the similarity of the associated items. But during training they do not make any distinction between items which are *highly relevant* and items which are only *partially* or *completely irrelevant* to a given query. For example, if a query is about 'how to cook a pizza', then videos which depict how to 'bake a pizza', 'cook pasta', or 'knead dough' are all treated the same way, although they can

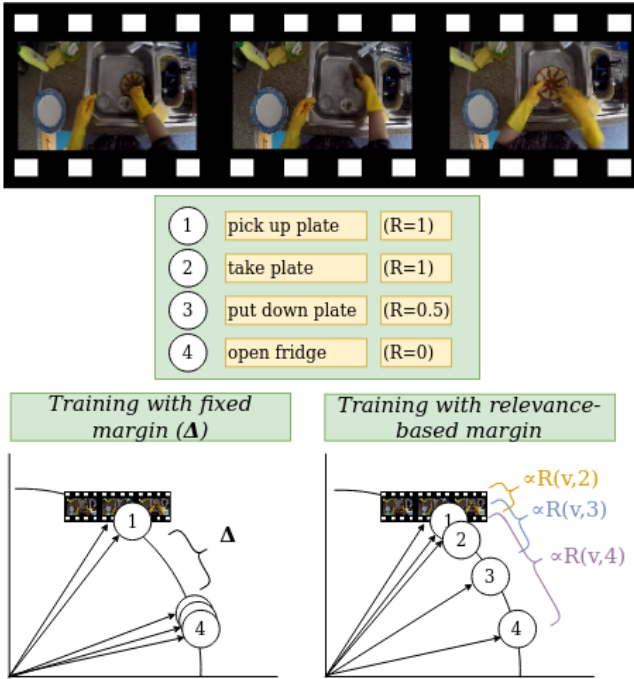


Figure 1: Training a model for text-video retrieval by employing a contrastive loss which uses a fixed margin Δ (lower left) treats semantically equivalent descriptions which do not appear as groundtruth pairs in the dataset as equally irrelevant. We propose to move away from such a paradigm and adopt a relevance-based margin (lower right), i.e. a margin which is proportional to the relevance \mathcal{R} (see Sec. 3.1).

be more or less semantically close to the query. Furthermore, one of the reasons which limited the usage of rank-aware metrics in video retrieval consists in visual-language datasets only providing the visual contents and textual annotations (obtained manually [Xu et al. 2016; Zhou et al. 2018] or automatically [Miech et al. 2019]). Due to the absence of relevance grades, rank-aware metrics (e.g. nDCG) are difficult to adopt. Recently, this problem was partially alleviated by the introduction of a relevance function [Damen et al. 2021a] which, to avoid a costly manual annotation step, is defined in terms of the captions already available in the dataset.

To give the model awareness of the semantical differences between items and queries during training, we free the margin from its stillness. Several solutions for non-fixed margins were proposed in previous literature, such as using multiple margins (e.g. [Cheng et al. 2016]) or adaptive solutions. In particular, [Semedo and Magalhães 2019] implemented a schedule for the margin value which gradually incorporates inter-category correlations and information about the structure of the embedding space. Recently, for video retrieval [He et al. 2021] proposed an adaptive margin proportional to the similarity of item and query as computed by multiple models. Differently from them, we propose to inject semantic knowledge into the training process by means of a relevance-based margin. To do so, we leverage the relevance function detailed in [Damen et al. 2021a], so that the margin is proportional to how relevant the item

is to the query, as illustrated in Fig. 1. By doing so, we effectively discard one hyper-parameter to tune. Moreover, even by performing an expensive search for it, the results are still suboptimal when compared to the proposed relevance-based margin. We give empirical evidence that the proposed technique makes it possible to easily improve the quality of the ranking lists, measured through Normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (mAP). We use three different and increasingly more complex models (MME from [Wray et al. 2019], JPoSE [Wray et al. 2019], and HGR [Chen et al. 2020b]) on two datasets (EPIC-Kitchens-100 [Damen et al. 2021a] and YouCook2 [Zhou et al. 2018]). Furthermore, we perform several ablations to study how it interacts with multiple video modalities (motion, appearance, audio) and with both cross-modal and within-modal losses.

We organize the paper as follows. In Section 2 we review related works, including vision and language tasks, main techniques and losses used to deal with text-video retrieval, and optimization of retrieval metrics such as the nDCG. Then, we formally describe the proposed technique in Sec. 3, in terms of the relevance function and how we apply it to a typical contrastive loss setting. In Sec. 4 we perform multiple experiments to prove the strength of the relevance-based margin. Finally, in Sec. 5 we conclude the paper.

2 RELATED WORKS

Vision and Language. In recent years, deep learning brought several advancements in multiple tasks dealing with vision and language, such as question answering [Anderson et al. 2018; Antol et al. 2015; Huang et al. 2020; Kim et al. 2020], retrieval [Chen et al. 2020b; Dong et al. 2021a; Lee et al. 2021; Zhang et al. 2020], and captioning [Dong et al. 2021b; Lei et al. 2020; Li et al. 2020a; Shi et al. 2021]. Given that vast amounts of data can be scraped from the web, many works perform a joint vision and language pretraining [Chen et al. 2020a; Li et al. 2020b; Sun et al. 2019; Zhou et al. 2021] by optimizing vision-text proxy tasks. Recently, a line of research uses natural language supervision such as captioning [Desai and Johnson 2021] or alignment [Jia et al. 2021] objectives to pretrain visual models. While in both cases they achieve competitive and state-of-the-art results on downstream tasks, these methods are data hungry and expensive to train, making them impractical from a computational point of view.

Text-Video Retrieval. Multiple techniques were proposed to learn a representation for the input data while capturing multimodal interactions. [Gabeur et al. 2020; Liu et al. 2019; Wang et al. 2021] explore multimodal fusion techniques to fuse all the information extracted from a video using multiple pretrained ‘experts’. While these methods focus on the addition of video-side information, a supervisory signal can also be obtained by looking with more detail at the text. [Chen et al. 2020b] create a semantic role graph of the caption and aligns to each node a learned representation of the clip-level descriptor. [Wray et al. 2019] extract verbs and nouns from the caption and uses them to learn Part-of-Speech-specific embedding spaces. [Patrick et al. 2020] introduce a generative cross-captioning task, using the batched videos as a support set. Recently [Croitoru et al. 2021] distil information from multiple pretrained text experts. A different trend involves heavy pretraining steps [Bain et al. 2021; Dzabraev et al. 2021; Lei et al. 2021; Liu et al.

2021], followed by finetuning for downstream tasks. Moreover, the addition of image-text datasets as part of the pretraining step, showed significant improvements when dealing with video-related tasks [Bain et al. 2021; Lei et al. 2021]. While these methods achieve impressive results, they rely heavily on the data, are expensive to train, and are not designed for the nature of the problem.

Due to the unavailability of groundtruth relevance values which can inform about the optimal ranking list to a given query, the video retrieval community focused on rank-unaware metrics such as the recall rates or the median rank. Contrastive losses greatly improve these metrics since they reduce the distance between the visual descriptor and the linguistic one and thus increase its similarity, making it possible to retrieve it before the negative descriptors. But multiple descriptions can be equally or partially relevant for the same video (and vice versa), thus more complex and rich metrics, such as the nDCG, are needed to accurately evaluate a retrieval system [Wray et al. 2021]. To do so, a way to determine how relevant an item is to a query must be available. To avoid the need for manual and costly annotation, [Damen et al. 2021a] proposes to use a relevance function defined in terms of the noun and verb classes present in the caption (more details in Sec. 3.1).

Learning a joint embedding space. Common approaches for text-video retrieval learn a joint embedding space by means of a contrastive loss [Hadsell et al. 2006; Schroff et al. 2015] which, during training, puts semantically similar items (e.g. a video and a caption describing its contents) closer in the embedding space, while dissimilar items are pushed away. While groundtruth associations (i.e. positive pairs, such as a video and its caption) are known from the dataset, the negative examples (such as a different video) have to be sampled, or ‘mined’, given that the amount of possible tuples scales exponentially with the dataset size, e.g. cubically with triplets. Multiple techniques have been proposed including: offline mining, which randomly samples a fixed number of tuples and repeats the process multiple times during training; online mining, which uses the negatives inside the mini-batch by considering all the non-groundtruth pairs, or only hard [Hermans et al. 2017; Xuan et al. 2020b] or semi-hard negatives [Schroff et al. 2015]. Recent research also found relevant signal while mining positive samples, e.g. easy [Xuan et al. 2020a] or hard positives [Hermans et al. 2017]. In our paper, we focus on triplets as they are a popular margin-based contrastive loss, but it can be extended to other techniques, e.g. to quadruplets [Chen et al. 2017]. Moreover, we experiment with two different and opposite techniques: offline mining with random sampling and online mining with hard negatives, and show the advantages of the relevance-based margin in both cases.

Margin in contrastive losses. Most of the approaches involving contrastive losses are based on maximum-margin losses (e.g. [Hadsell et al. 2006]). Although the margin is usually fixed, variable or adaptive solutions for it have been explored in different fields. For person re-identification, [Cheng et al. 2016] suggest using two different (but fixed) margins for inter- and intra-class constraints, whereas [Zhang et al. 2019] propose to monotonically increase the margin during the training process. [Hu et al. 2018] use a ‘soft margin’ to improve recommender systems, that is they remove the fixed margin and use (a soft version of) the distance between positive and negative pairs as the loss. [Li et al. 2020c] augment the bidirectional contrastive loss by also summing the margin to the loss objective,

to optimize it during the training process. For text-image retrieval, [Semedo and Magalhães 2019] propose a scheduled adaptive margin which starts from a fixed value and gradually changes during the training process both to integrate inter-category similarity-based correlations and to preserve the category clusters formed during the initial phases of the training. Recently, for cross-modal video retrieval [He et al. 2021] proposed an adaptive margin proportional to the similarity of the representations computed for the negative pair, both in terms of ‘static’ (pretrained, frozen) models, which provide initial supervision, and ‘dynamic’ (trained with the task) models, which provide supervision in later stages of the training. Differently from all these works, we propose a margin which is proportional to the relevance value of the queries involved in the triplet, effectively using the semantic knowledge during training.

Optimization of nDCG. Considering that visual-textual datasets usually lack relevance grades, rank-unaware metrics are one of the preferred ways to measure progress in the video retrieval community. Yet given a video, multiple captions can be used to describe its contents. To capture the difference in the ranking list when binary relevance (i.e. a caption is either relevant or irrelevant to a video) is considered, mAP is preferred to the recall rates. Furthermore, finer-grained relevance grades could be also available (i.e. a caption can be relevant to a video to some degree), in which case the DCG (or its normalized version, the nDCG) is chosen. But, optimizing these metrics during training clashes with gradient-based optimization methods because ranks are not differentiable with respect to the learnable parameters, e.g. the nDCG of a list of items to a given query is normalized using the optimal ranking list, which is computed by *sorting* with respect to the relevance values.

Surrogate losses are used to partially address this problem, which can be categorized into: pointwise (e.g. regression loss [Cossock and Zhang 2008]), which compare predicted and optimal rank of one item at a time; pairwise (e.g. RankNet [Burges et al. 2005]), which deal with pairs of items and relative ordering; listwise approaches (e.g. LambdaRank [Burges et al. 2006]), which work on full list of items. Note that the triplet loss [Schroff et al. 2015] can be seen as a ‘triplet-wise’ surrogate loss. Since these surrogate losses are loosely connected to downstream metrics, there is also an active research field which directly optimizes retrieval metrics by deriving a relaxation of the sorting operator which has well-defined gradients, e.g. [Blondel et al. 2020; Cuturi et al. 2019; Grover et al. 2018].

Considering its widespread usage for video retrieval, we consider the triplet loss an optimal candidate for our relevance-based margin, and show it can lead to higher quality ranking lists.

3 RELEVANCE-BASED MARGIN

In Sec. 3.1 we define the relevance function \mathcal{R} and the metrics used during evaluation. In Sec. 3.2 we describe how we change the margin in the contrastive loss to make it dependent on \mathcal{R} . Finally, Sec. 3.3 details the three methods on which we test our technique.

3.1 Semantic classes and relevance

Given a video clip, multiple natural language descriptions may fully capture its visual contents, and vice versa. Hence, if a user looks for videos about ‘cooking a pizza’, an intelligent video retrieval system should retrieve all the videos which show how to cook a pizza, and

show them all before (i.e. rank them higher than) those that show the baking of a ‘focaccia’. Similarly, videos about ‘fried potatoes’ should be ranked even lower, given how dissimilar they are when compared to the user query. As a consequence, the automatic evaluation of the quality of a ranking list requires a function which considers ‘focaccia’ more relevant than ‘potatoes’ when compared with ‘pizza’, as well as the cooking technique (‘bake’ versus ‘fry’). To avoid the need for costly manual annotation which requires human assessments using a predefined set of grades, [Damen et al. 2021a] introduces a relevance function \mathcal{R} defined as:

$$\mathcal{R}(x_i, x_j) = \frac{1}{2} \left(\frac{|x_i^V \cap x_j^V|}{|x_i^V \cup x_j^V|} + \frac{|x_i^N \cap x_j^N|}{|x_i^N \cup x_j^N|} \right) \quad (1)$$

where x_k^V and x_k^N denote the sets of verb and noun classes found in the k -th caption. This can be extended to videos by considering the associated description. By defining the relevance as in Eq. 1, x_i is highly relevant to x_j if they share similar noun and verb classes. We refer to ‘classes’ because we do not want to consider synonyms (e.g. ‘pick up’ and ‘take’, or ‘drop’ and ‘put down’) as different items which need to be separated, hence each class will contain tokens with a similar meaning. In some datasets, this class knowledge may be already available, but several other datasets do not provide it. To automatically compute them, a pipeline made of a PoS-tagger (e.g. with spaCy), followed by WordNet [Miller 1995] and the Lesk algorithm [Lesk 1986] can be used, as in [Wray et al. 2021].

To evaluate a video retrieval system, we use two metrics which are commonly used in Information Retrieval, which are the Mean Average Precision (mAP [Baeza-Yates et al. 1999]) and the Normalized Discounted Cumulative Gain (nDCG [Järvelin and Kekäläinen 2002]), as recently proposed in [Wray et al. 2021]. The mAP is defined as the mean of the Average Precision (AP) with respect to all the queries. For a given query q , AP can be defined as:

$$AP(q) = \frac{\sum_{k=1}^N P(k) \cdot r(k)}{N_r} \quad (2)$$

where N is the number of items (both relevant and irrelevant) in the ranking list, $P(k)$ is the Precision at k [Baeza-Yates et al. 1999], $r(k)$ is an indicator function which tells whether the k -th item is relevant or not, and N_r is the total number of relevant items. The mAP allows to grasp with a single number the area under the Precision-Recall curve. But this metric requires binary relevance values, thereby requiring the introduction of a threshold below which items are considered irrelevant (and relevant otherwise). For mAP, we consider k to be relevant to q only when $\mathcal{R}(q, x_k) = 1$ as is done in [Damen et al. 2021a] (hence, for mAP $N_r = |\{x_i \mid \mathcal{R}(q, x_i) = 1\}|$). On the other hand, nDCG makes use of non-binary relevance values, allowing it to grasp finer details (and errors) of the ranking list. Given a query q and a list of items $K = \{x_i\}$, it is defined as

$$nDCG(q, K) = \frac{DCG(q, K)}{IDCG(q, K)} \quad (3)$$

where IDCG is the optimal DCG value obtained when the ranking list follows a descending order of relevance values. We define DCG as in [Damen et al. 2021a; Järvelin and Kekäläinen 2002]:

$$DCG(q, K) = \sum_{k=1}^{N_r} \frac{\mathcal{R}(q, x_k)}{\log_2(k+1)} \quad (4)$$

where x_k is the k -th item in the list K , and we only consider the first N_r items in the ranking list. Note that $N_r = |\{x_i \mid \mathcal{R}(q, x_i) > 0\}|$.

3.2 Contrastive loss with relevance-based margin

To learn a joint text-video embedding space, various contrastive (or ranking) losses have been proposed (see Sec. 2). In our work we consider a contrastive term based on the triplet loss defined as:

$$\mathcal{L} = [m + s(a, n) - s(a, p)]_+ \quad (5)$$

where $[\cdot]_+ = \max(0, \cdot)$, m is interpreted as a separation margin, $s(\cdot, \cdot)$ is a similarity metric (e.g. cosine similarity), whereas a , n , and p represent respectively the embedding of the *anchor*, *negative*, and *positive* item. Eq. 5 provides a positive loss when the margin m between the positive pair (a, p) and the negative one (a, n) is violated, i.e. $s(a, p) - s(a, n) < m$. The loss may be cross-modal, i.e. n, p from one modality (e.g. video) and a from the opposite one (e.g. text), or within-modal, i.e. a, p, n are all from the same modality. Furthermore, the optimal m is not known beforehand and should be treated as an hyper-parameter which can affect the performance. Thus, it should be tuned on the validation set.

During training, all the items which are not from the positive pair (a, p) are pushed away until they are separated by a margin of m , as shown in Fig. 1. Although effective and widely used in the literature, Eq. 5 ignores that multiple items may be completely or partially relevant to the same query, and treats all the items which are not from the groundtruth pair as equally irrelevant. Thus the retrieval system might not be able to distinguish between the many relevance levels which can exist between an item and a query.

To address this, we propose a relevance-based margin instead of a fixed margin. In our work, we aim at defining m in terms of the relevance function \mathcal{R} . In particular, we update Eq. 5 as follows:

$$\mathcal{L} = [\Delta_{a,p,n} + s(a, n) - s(a, p)]_+ \quad (6)$$

where:

$$\begin{aligned} \Delta_{a,p,n} &= R(a, p) - R(a, n) \\ &= 1 - R(a, n) \end{aligned} \quad (7)$$

since we consider the groundtruth pair to be maximally relevant, i.e. $R(a, p) = 1$. The relevance-based margin keeps \mathcal{L} positive until $s(a, p)$ and $s(a, n)$ are separated by a margin which is proportional to their relevance values, thus separating irrelevant items more than those which have a positive relevance. This is illustrated in Fig. 1 on the lower right. Note that this term is not bound to the network state and can thus be applied both to offline and online mining techniques.

3.3 Methods

Given a dataset $D = \{(v_i, q_i)\}$ of video-caption pairs, we strive to learn optimal weights for two embedding functions $f : \mathbb{R}^{f_v} \rightarrow \mathbb{R}^d$ and $g : \mathbb{R}^{f_q} \rightarrow \mathbb{R}^d$ such that $f(v_i)$ and $g(q_i)$ are close in the d -dimensional joint embedding space. Here f_v and f_q represent the dimension of the video and caption descriptors. To parameterize f and g we consider the following methods: **MME** is a baseline from [Wray et al. 2019] which learns one embedding function for each of the two modalities, video and text. In **JPoSE** [Wray et al. 2019], the captions are processed in order to obtain a single sentence-level

descriptor and multiple descriptors restricted to specific Part-of-Speech (PoS) tags, e.g. nouns and verbs. Then, two functions are learned for each of these embedding spaces using both cross-modal and intra-modal contrastive terms for the sentence-level, as well as for the PoS-level. HGR [Chen et al. 2020b] structures the learning at multiple levels (global event, local actions, and local entities) which are obtained by computing a semantic role graph for each of the captions. Then a graph convolutional network is used to learn compositional semantics of the caption based on the local components, i.e. full sentence, verbs, and noun phrases.

We choose these three methods because they provide incrementally structured approaches to deal with video and language data, starting from a simpler MLP-based network to a graph-based approach. Moreover, JPoSE represents the state-of-the-art for EPIC-Kitchens-100 (measured through nDCG and mAP), which is the main dataset under consideration. Finally, by selecting them we can validate our approach on both offline (MME and JPoSE) and online (HGR) mining techniques. We thus proceed to show the generality and effectiveness of the proposed relevance-based margin by empirically validating on two different datasets.

4 EXPERIMENTS

After the introduction of the experimental setting in Sec. 4.1, we show in Sec. 4.2 how the proposed relevance-based margin helps to achieve better nDCG and mAP on EPIC-Kitchens-100 and YouCook2. Then, in Sec. 4.3 we perform several ablation studies. First we show that even by carefully tuning the fixed margin, the proposed technique consistently achieves better results without the need to tune it. Secondly, we also evaluate its robustness by ablating the loss function and the modalities used in JPoSE. Finally in Sec. 4.4 we analyze the distribution of the margin values during training and some video-to-text examples from the testing set.

4.1 Experimental setting

Datasets. We focus our experimental setting on two challenging video and language datasets: the recently released EPIC-Kitchens-100 [Damen et al. 2021a] and YouCook2 [Zhou et al. 2018]. For the retrieval challenge, the former comprises 67217 egocentric clips for training and 9668 for evaluation. It is also the largest dataset for video retrieval in the egocentric setting. Moreover, it also provides semantic annotations for each of the captions, by covering 300 noun and 97 verb classes. The latter provides a lower amount of training clips (10337) but still offers a challenging evaluation set with 3492 clips. While semantic annotations are not available for YouCook2 they can be computed using WordNet and the Lesk algorithm, as described in Sec. 3.1. Furthermore, as both EPIC-Kitchens-100 and YouCook2 share the kitchens domain, the class knowledge of the former can also be used for the latter [Wray et al. 2021].

Implementation details. For EPIC-Kitchens-100 we use the TBN [Kazakos et al. 2019] features from the dataset provider comprising of 25 uniformly sampled RGB, flow, and audio feature vectors for each clip. For YouCook2 we use ImageNet-pretrained ResNet-152 features from the VALUE benchmark [Li et al. 2021]. For the three methods we use the open source codebases provided in the respective papers and follow their hyper-parameter setting. We release our code and models on GitHub to support reproducibility.

Method	rel- Δ	nDCG	mAP
MME		48.5	38.5
MME	✓	49.6 \uparrow 1.1	39.2 \uparrow 0.7
JPoSE		53.5	44.0
JPoSE	✓	56.2 \uparrow 2.7	45.8 \uparrow 1.8
HGR		32.2	36.0
HGR	✓	50.2 \uparrow 18	45.6 \uparrow 9.6

Table 1: nDCG and mAP results on EPIC-Kitchens-100 with three different methods, using TBN (RGB, Flow, Audio) features. We report in bold the best results (and underline the second best). With “ \uparrow X” we represent an improvement of X when compared to the above result.

4.2 Relevance-based margin results

EPIC-Kitchens-100. To validate the effectiveness of the proposed relevance-based margin, we explore three methods (MME, JPoSE, and HGR as described in Sec. 3.3) on EPIC-Kitchens-100. In Tab. 1 we report nDCG and mAP values, averaged between text-to-video and video-to-text. In all three cases, we observe a large improvement in both metrics, showing that the relevance-based margin works on very different models. It also works well with both offline mining with randomly sampled triplets (for MME and JPoSE), and online mining with hard negatives (for HGR): by using the relevance-based margin, MME gains +1.1 nDCG and +0.7 mAP, JPoSE +2.7 nDCG and +1.8 mAP, and finally HGR obtains +18 nDCG and +9.6 mAP. Such a large improvement is possibly due to how the triplets are sampled: in JPoSE, the negatives do not share the verb class of the anchor, leading to a relevance lower than 0.5; but, there is not such a guarantee in HGR, since batches are formed randomly. Hence, by employing a relevance-based margin in HGR we automatically deal with situations in which the negatives have a considerable relevance and adapt the margin accordingly. Finally, in App. A we report the public leaderboard for the retrieval challenge, confirming the improvement we observe over current state-of-the-art methods.

YouCook2. In the previous experiment we used the class knowledge which accompanies the dataset. But, by computing synsets knowledge in a similar way to what is done in EPIC-Kitchens-100, the proposed relevance-based margin can still successfully help the training process. This setting poses two additional challenges: first of all, in EPIC-Kitchens-100 most of the captions follow a precise structure, i.e. they contain a verb and a noun, which is not the case when dealing with other datasets, where free-form descriptions are often adopted. This may make it more difficult for the PoS-tagger to correctly tag the words. Secondly, there may be words which are put in the wrong category by WordNet.

For this dataset, we use the same class knowledge used in EPIC-Kitchens-100, as it transfers well across both datasets since they share the cooking domain [Wray et al. 2021], and for words which do not appear in any class, a new singleton class is created.

In Tab. 2 we report the nDCG and mAP values obtained with MME, JPoSE, and HGR. From the table, one can see that even in this different setting the relevance-based margin is able to provide useful information to the model. For example, the addition of the proposed technique in HGR leads to a gain of +5.5 nDCG and +3.1 mAP when compared to the results obtained with a fixed margin.

Method	rel- Δ	nDCG	mAP
MME	✓	46.9	19.3
		47.3 \uparrow 0.4	19.5 \uparrow 0.2
JPoSE	✓	49.6	20.5
		50.4 \uparrow 0.8	21.5 \uparrow 1.0
HGR	✓	41.0	23.0
		46.5 \uparrow 5.5	26.1 \uparrow 3.1

Table 2: nDCG and mAP using MME, JPoSE, and HGR on YouCook2. We use ResNet-152 (pretrained on ImageNet) features from the VALUE benchmark [Li et al. 2021].

4.3 Ablation studies

We perform the ablation studies on EPIC-Kitchens-100 using JPoSE.

Varying the fixed margin. In Sec. 4.2 we show that the proposed relevance-based margin leads to improved nDCG and mAP on both EPIC-Kitchens-100 and YouCook2. But, what if one would only need to carefully tune the fixed margin to obtain similar results? To answer to this question, we focus on JPoSE and vary the fixed margin Δ in $\{0.1, 0.2, \dots, 1.5\}$ (default value used in JPoSE is 1.0). We keep the rest of the hyper-parameter setting as in [Damen et al. 2021a; Wray et al. 2019] and use the officially provided TBN features. We plot in Fig. 2 nDCG, mAP, average R@1 for each of the tested margins. While small margins lead to worse results overall, it can be seen that increasing the margin does not improve significantly neither the nDCG nor the mAP. Moreover, the recall rates are affected only marginally as well. When compared to the performance shown by the adoption of the relevance-based margin, it can be observed that our technique manages to achieve higher nDCG and mAP values, while also keeping similar recall rates (on average, 6.3% R@1). Finally, it is worth noticing that by using the relevance-based margin we are released from the margin hyper-parameter: this is also of practical importance, because by using a fixed margin its optimal value is not known in a testing scenario, hence one would also need to perform an expensive search on the validation set in order to achieve better performance.

Losses ablation. A peculiarity of JPoSE is that it uses multiple contrastive loss terms to learn both global- and PoS-restricted joint embedding spaces. To do so, the authors employ a global loss and a PoS-level loss, both in the cross- and within-modality settings. We perform an ablation study in Tab. 3 to give evidence that the relevance-based margin can be helpful even when restricting the amount of loss terms used. Note that when applying the technique to the PoS-level terms (e.g. verbs) we consider the term for the opposite PoS (e.g. nouns) in Eq. 1 to be 1. As shown in Tab. 3, the adoption of the relevance-based margin leads to an improvement of +1.6 nDCG and +1.2 mAP when using only the cross-modal global-level loss terms, whereas +2.8 nDCG and +1.9 mAP are gained when also adding the cross-modal PoS-level terms.

Modalities ablation. For EPIC-Kitchens-100 we have RGB, flow, and audio features. To show that the improvements obtained when applying the relevance-based margin are not due to the model accessing multiple modalities related to the video, we perform another ablation in Tab. 4 by considering RGB-only and RGB+flow features. In both cases the proposed technique shows its usefulness. In particular, by employing the relevance-based margin we observe

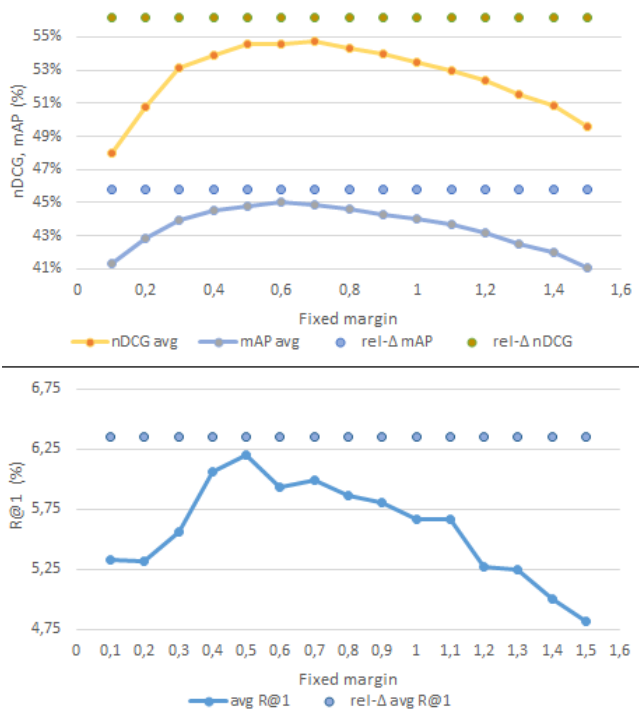


Figure 2: Using JPoSE on EPIC-Kitchens-100, we show how changing the fixed margin in the loss function affects the performance, measured through nDCG and mAP in the upper figure, and average R@1 in the lower one. For reference, we also plot disconnected dots to show the performance when we use the proposed relevance-based margin. Notice that the optimal fixed-margin hyper-parameter would not be known in a testing scenario; it would need to be estimated through an expensive hyper-parameter search on a validation set.

rel- Δ	cross-glob	within-glob+PoS	nDCG	mAP
	✓		53.1	43.3
✓	✓		54.7 \uparrow 1.6	44.5 \uparrow 1.2
	✓	✓	53.4	43.7
✓	✓	✓	56.2 \uparrow 2.8	45.6 \uparrow 1.9
	✓	✓	53.5	44.0
✓	✓	✓	56.2 \uparrow 2.7	45.8 \uparrow 1.8

Table 3: nDCG and mAP using JPoSE on EPIC-Kitchens-100. During training, JPoSE considers both cross- and within-modality contrastive losses, both at sentence- and PoS-level. Applying the relevance-based margin helps at each level.

+1.6 nDCG and +1.6 mAP when using RGB-only, +2.9 nDCG and +1.8 mAP when using both RGB and flow, and +2.7 nDCG and +1.8 mAP when adopting all the three modalities.

Modalities	rel- Δ	nDCG	mAP
RGB	✓	36.8	28.8
		38.4 \uparrow 1.6	30.4 \uparrow 1.6
RGB+Flow	✓	49.6	41.0
		52.5 \uparrow 2.9	42.8 \uparrow 1.8
RGB+Flow+Audio	✓	53.5	44.0
		56.2 \uparrow 2.7	45.8 \uparrow 1.8

Table 4: TBN offers RGB, flow, and audio features. The proposed relevance-based margin interacts with each modality in an incremental way. We use JPoSE on EPIC-Kitchens-100.

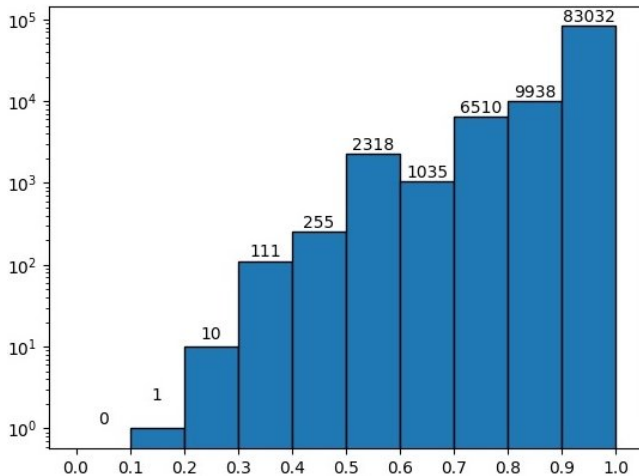


Figure 3: Log-scale distribution of the margins used during training. Over each bin we report the frequency. Numbers refer to one epoch, with 10 triplets sampled for each example (e.g. with 10337 examples for YouCook2, we end up with around 103k triplets). Although a great part of the triplets are separated with the highest margin (i.e. lowest relevance), around 20k triplets are distanced by various margin values.

4.4 Qualitative analysis

First of all, the proposed technique leads to variable margins, therefore the distribution of the values may help explaining why we observe such a positive influence on the final performance. In Fig. 3 we plot the frequencies of the margins (with bins of size 0.1) observed during the training of JPoSE on YouCook2, where for each of the training examples 10 triplets are sampled. It can be seen that a great part of the margins used are in the final bin (between 0.9 and 1.0), for which the relevance is quite low since the margin is computed as $\Delta_{a,p,n} = 1 - \mathcal{R}(a, n)$ (see Eq. 7). In these cases, the margin will be similar to the default case of JPoSE, i.e. 1.0. Yet, around 20% of the training triplets end up having smaller margins. In these situations, the varying margins help the model achieve better performance by providing a semantic supervision on the structure of the embedding space, since the relevant items are kept at a distance which is proportional to the relevance.

Secondly, in Fig. 4 we visualize a few video-to-text examples from the testing set, by plotting for each of them the relevance

values of each caption in both the full ranking list and the top 50 retrieved captions. By plotting the full ranking list, it is possible to see that the relevance-based margin helps improving the nDCG, as relevant captions are retrieved first. This can also be seen in the top 50 of Fig. 4.a, 4.b, and 4.c where with the relevance-based margin no irrelevant captions are retrieved and, especially in Fig. 4.c, the ranking is almost ideal. Yet we can still find examples where the proposed technique fails to achieve the expected improvements. In Fig. 4.d, using the relevance-based margin a few irrelevant captions are retrieved, such as ‘take container’ and ‘take milk container’. This behavior is likely related to the fact that during training captions like ‘close container’ and ‘close milk container’ are relevant (0.5) for a video depicting the action ‘close fridge’, since they share the same verb class. This leads to an increase in the similarity of the respective descriptors. Hence, during inference, also captions like ‘take container’ and ‘take milk container’ might have a significant similarity with the video descriptor of ‘close fridge’. Further qualitative analysis is available in Appendix B.

5 CONCLUSIONS

Learning a joint embedding space using a margin-based contrastive loss is the dominant approach to deal with text-video retrieval. In the literature it is shown that by using such a framework, competitive performance on rank-unaware metrics, such as the recall rates, can be obtained. Yet, rank-aware metrics, such as the nDCG, need to be taken into account, as multiple descriptions can have numerous levels of relevance to a given query [Wray et al. 2021]. In this work, we proposed to move away from the fixed margin which is typically used in such a framework, and introduced a relevance-based margin. In particular, we adopted the proposed technique into three increasingly more complex models on two datasets and gave empirical evidence that we can easily improve the performance measured through nDCG and mAP. Moreover, we showed that even by performing an expensive search of the fixed margin hyper-parameter, it does not reach the same performance as when using the relevance-based margin. Furthermore, the proposed technique can also have a positive impact on video retrieval applications as not needing to tune the margin can lead to less GPU hours required to fully train the model. Finally, we focused our work on text-video retrieval, but the relevance-based margin can be easily extended to other domains where similar margin-based ranking losses are used, e.g. in image retrieval [Zhang et al. 2020]. Moreover, we showed the effectiveness of the proposed approach by applying it to loss functions where the margin is explicitly defined and used to separate positive and negative pairs, e.g. [Chen et al. 2017; Schroff et al. 2015]. Yet, there are also popular loss functions which do not make use of it, such as NCE [Gutmann and Hyvärinen 2010] and MIL-NCE [Miech et al. 2020]. Future work is required to adapt the relevance-based margin to non-margin based loss functions.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from Amazon AWS Machine Learning Research Awards (MLRA) and NVIDIA AI Technology Centre (NVAITC), EMEA. We acknowledge the CINECA award under the ISCRA initiative, which provided computing resources for this work. This work has been partially supported by the Spanish

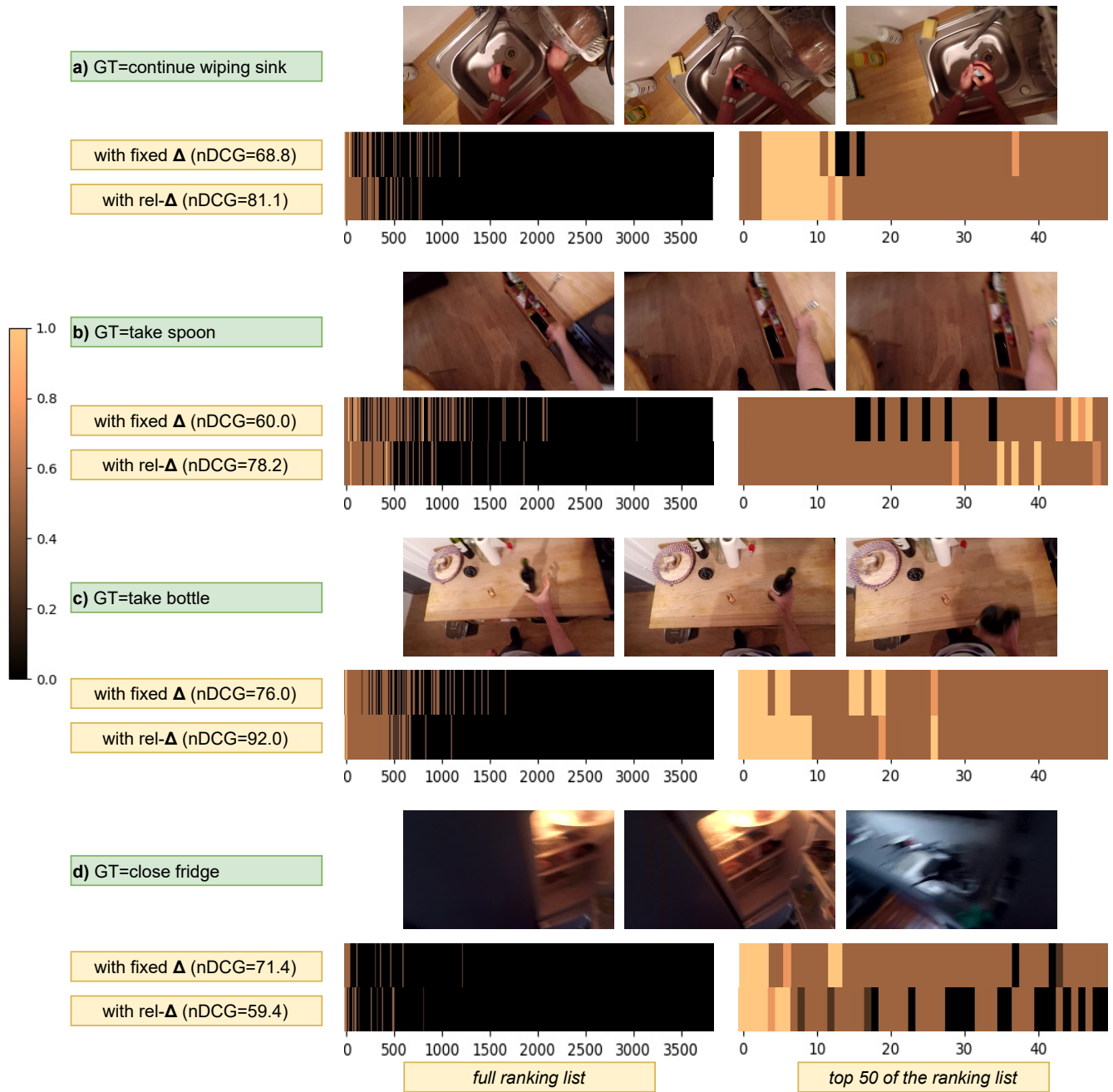


Figure 4: Video-to-text qualitative results on EPIC-Kitchens-100 testing set using JPoSE. For each of the examples we show a few frames and the groundtruth (GT) caption, and we plot both the full ranking list and the top 50 retrieved captions when adopting the fixed margin and then the relevance-based margin. On the left we also visualize the color bar which is used for the relevance (light colors mean high relevance, dark colors low relevance). In particular, Figures a, b, and c are success cases, whereas Figure d represents a failure case.

project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

- Stanislaw Antol, Aishwarya Agrawal, Jiaseen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *ICCV* (2021).
- Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. 2020. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*. PMLR, 950–959.
- Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with non-smooth cost functions. *Advances in neural information processing systems* 19 (2006), 193–200.
- Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. 89–96.
- Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020b. Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 403–412.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020a. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
- De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1335–1344.
- David Cossack and Tong Zhang. 2008. Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory* 54, 11 (2008), 5140–5154.
- Ioana Croitoru, Simion-Vlad Bogolin, Marius Lordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. 2021. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11583–11593.
- Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. 2019. Differentiable ranks and sorting using optimal transport. *NeurIPS* (2019).
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2021a. Rescaling egocentric vision. *IJCV* (2021).
- Dima Damen, Adriano Fragomeni, Jonathan Munro, Toby Perrett, Daniel Whetam, Michael Wray, Antonino Furnari, Giovanni Maria Farinella, and Davide Moltisanti. 2021b. *EPIC-KITCHENS-100- 2021 Challenges Report*. Technical Report. University of Bristol.
- Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11162–11173.
- Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021a. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- Xinzhong Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. 2021b. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2615–2624.
- Maksim Dzabaraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3354–3363.
- Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal Transformer for Video Retrieval. In *Proceedings of the IEEE European Conference on Computer Vision*. Springer.
- Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. 2018. Stochastic Optimization of Sorting Networks via Continuous Relaxations. In *International Conference on Learning Representations*.
- Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.
- Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, Vol. 2. IEEE, 1735–1742.
- Feng He, Qi Wang, Zhifan Feng, Wenbin Jiang, Yajuan Lü, Yong Zhu, and Xiao Tan. 2021. Improving Video Retrieval by Adaptive Margin. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1359–1368.
- Alexander Hermans, Lucas Beyler, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
- Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7258–7267.
- Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Minghui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11021–11028. Issue 07.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML* (2021).
- Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2019. Epicfusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5492–5501.
- Junyeon Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. 2020. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10106–10115.
- Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 802–812.
- Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7331–7341.
- Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2603–2614.
- Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*. 24–26.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200* (2020).
- Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. 2021. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.
- Mingming Li, Shuai Zhang, Fuqing Zhu, Wanhui Qian, Liangjun Zang, Jizhong Han, and Songlin Hu. 2020c. Symmetric metric learning with adaptive margin for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4634–4641.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.
- Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval. *ICCV* (2021).
- Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *BMVC* (2019).
- Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9879–9889.
- 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.
- George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metz, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824* (2020).
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
- David Semedo and João Magalhães. 2019. Cross-Modal Subspace Learning with Scheduled Adaptive Margin Constraints. In *Proceedings of the 27th ACM International Conference on Multimedia*. 75–83.

- Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing Descriptive Image Captioning with Natural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 269–277.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 7464–7473.
- Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021. T2vLad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5079–5088.
- Michael Wray, Hazel Doughty, and Dima Damen. 2021. On Semantic Similarity in Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3650–3660.
- Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*. 450–459.
- 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 conference on computer vision and pattern recognition*. 5288–5296.
- Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020b. Hard negative examples are hard, but useful. In *European Conference on Computer Vision*. 126–142.
- Hong Xuan, Abby Stylianou, and Robert Pless. 2020a. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2474–2482.
- Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3536–3545.
- Yingying Zhang, Qiaoyong Zhong, Liang Ma, Di Xie, and Shiliang Pu. 2019. Learning incremental triplet margin for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9243–9250.
- Luowei Zhou, Jingjing Liu, Yu Cheng, Zhe Gan, and Lei Zhang. 2021. CUPID: Adaptive Curation of Pre-training Data for Video-and-Language Representation Learning. *arXiv preprint arXiv:2104.00285* (2021).
- Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*. 7590–7598. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344>

A COMPARISON WITH THE EPIC-KITCHENS-100 CHALLENGE LEADERBOARD

The release of the EPIC-Kitchens-100 dataset [Damen et al. 2021a] was accompanied by a public challenge for the multi-instance retrieval problem (alongside other challenges, e.g. for Action Recognition). To further prove the results we show in Section 4, we took part into the challenge by employing the proposed relevance-based margin on the JPoSE method [Wray et al. 2019] (see Section 3). We show the results of both the participants at the time of submission and those that took part into the previous challenge (which ended in November 2021) in Figure 5. The previous best result was obtained by Hao et al. (more details in the technical report [Damen et al. 2021b]), which achieved on average 44.23% mAP and 53.56% nDCG. As can be seen, we achieve 45.86% mAP (+1.63%) and 56.21% nDCG (+2.65%).

B QUALITATIVE ANALYSIS

We further analyze the effectiveness of the proposed technique from a qualitative point of view. To do so, we select three types of information. First of all, we pick a caption and compute its embedding (q) , pick the corresponding video descriptor (v) , and compute their similarity $s(v, q)$ through dot product. Then, we look for 10 similar captions (i.e. different captions which either share the noun or the verb class), pick the corresponding video descriptors indexed by V_+ , and compute an aggregated similarity value $s(v_+, q) = \frac{1}{10} \sum_{v_i \in V_+} s(v_i, q)$. Finally, we also randomly select 10

dissimilar captions (i.e. sharing neither the verb nor the first noun class), pick their video descriptors, and compute $s(v_-, q)$. We compare the results using JPoSE on the testing set of EPIC-Kitchens-100, and report several examples in Figure 6. In Figures 6.a and 6.b the usage of a fixed margin leads to a far too high similarity of the videos in V_+ with the query q when compared to its groundtruth video descriptor v , which hurts both nDCG, mAP, and the recall rates. In Figures 6.c and 6.d the videos in V_- and those in V_+ are not properly separated, hence wrongly giving the model the impression that they are similarly relevant to the query q . In all these cases, adopting a relevance-based margin is a successful strategy to correct these wrong predictions, leading to a more robust model.

Test set													2022 Challenge (currently open)		
#	User	Entries	Date of Last Entry	SLS			Mean Average Precision (%)			Normalised Discounted Cumulative Gain (%)					
				PT ▲	TL ▲	TD ▲	Avg. ▲	T2V ▲	V2T ▲	Avg. ▲	T2V ▲	V2T ▲			
1	afalcon	1	01/28/22	2.00 (1)	3.00 (1)	3.00 (1)	45.86 (1)	40.36 (1)	51.36 (1)	56.21 (1)	54.23 (1)	58.19 (1)			
2	MI-MM	1	12/10/21	2.00 (1)	3.00 (1)	3.00 (1)	27.58 (2)	23.08 (2)	32.09 (2)	42.10 (2)	40.48 (2)	43.72 (2)			

Test set													2021 Challenge (closed)		
#	User	Entries	Date of Last Entry	Team Name	SLS			mean Average Precision (%)			normalised Discounted Cumulative Gain (%)				
					PT ▲	TL ▲	TD ▲	Avg. ▲	T2V ▲	V2T ▲	Avg. ▲	T2V ▲	V2T ▲		
1	haoxiaoshuai	6	04/08/21	IIE_MRG	2.0 (1)	3.0 (1)	3.0 (1)	44.23 (1)	38.49 (1)	49.96 (1)	53.56 (1)	51.83 (1)	55.28 (2)		
2	JPoSE	3	01/07/21		2.0 (1)	3.0 (1)	3.0 (1)	44.01 (2)	38.11 (2)	49.91 (2)	53.53 (2)	51.55 (2)	55.51 (1)		
3	MLP	6	01/06/21		2.0 (1)	3.0 (1)	3.0 (1)	38.49 (3)	33.99 (3)	42.99 (3)	48.49 (3)	46.92 (3)	50.05 (3)		
4	MI-MM	4	05/06/21		2.0 (1)	3.0 (1)	3.0 (1)	29.21 (4)	23.60 (4)	34.83 (4)	44.79 (4)	42.40 (4)	47.18 (4)		

Figure 5: We report the public leaderboard for the EPIC-Kitchens-100 Challenge at time of submission (below), and also the leaderboard for the previous challenge which ended in November 2021 (above). It can be seen that we achieve around +1.6% mAP and +2.6% nDCG over the previous best results, achieved by Hao et al. (details in the technical report [Damen et al. 2021b]).

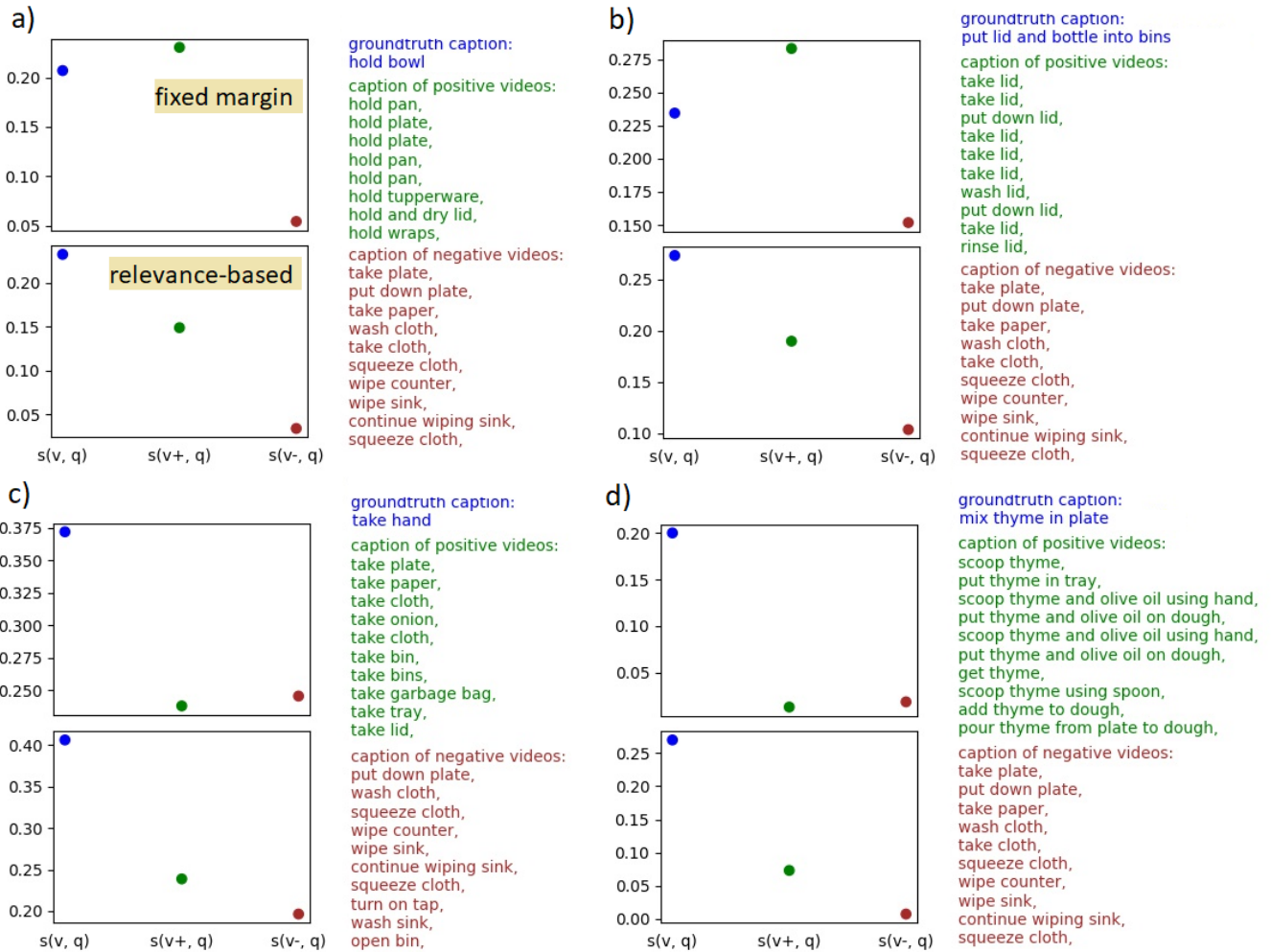


Figure 6: Using JPoSE, we compute a similarity score $s(v, q)$ for the groundtruth pair (colored in blue), $s(v+, q)$ for videos with similar captions (colored in green), and $s(v-, q)$ for videos with dissimilar captions (colored in brown). Note that, when selecting V_+ , for the examples on the left we change the noun class, whereas on the right we change the verb class. See Sec. B for more details. The captions of the videos used are reported on the right. Each of the four examples are taken from EPIC-Kitchens-100 testing set and for each of them we report first what happens with fixed margin, then with the proposed relevance-based margin.