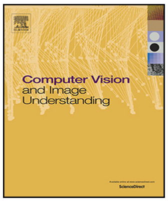
Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Improving semantic video retrieval models by training with a relevance-aware online mining strategy

Alex Falcon [a,b,*], Giuseppe Serra [a], Oswald Lanz [c]

[a] University of Udine, Via delle Scienze, 206, Udine 33100, Italy
[b] Fondazione Bruno Kessler, Via Sommarive, 18, Povo 38123, Italy
[c] Free University of Bozen-Bolzano, Piazza Domenicani 3, Bolzano 39100, Italy

ARTICLE INFO

Communicated by Yu-Chiang Frank Wang

Dataset link: https://github.com/epic-kitchens/C5-Multi-Instance-Retrieval, https://github.com/ninatu/everything_at_once, https://github.com/aranciokov/ranp/

Keywords:
Video retrieval
Cross-modal retrieval
Contrastive loss
Negative samples mining
Positive samples mining

ABSTRACT

To retrieve a video via a multimedia search engine, a textual query is usually created by the user and then used to perform the search. Recent state-of-the-art cross-modal retrieval methods learn a joint text–video embedding space by using contrastive loss functions, which maximize the similarity of *positive* pairs while decreasing that of the *negative* pairs. Although the choice of these pairs is fundamental for the construction of the joint embedding space, the selection procedure is usually driven by the relationships found within the dataset: a positive pair is commonly formed by a video and its own caption, whereas unrelated video-caption pairs represent the negative ones. We hypothesize that this choice results in a retrieval system with limited semantics understanding, as the standard training procedure requires the system to discriminate between groundtruth and negative even though there is no difference in their semantics. Therefore, differently from the previous approaches, in this paper we propose a novel strategy for the selection of both positive and negative pairs which takes into account both the annotations and the semantic contents of the captions. By doing so, the selected negatives do not share semantic concepts with the positive pair anymore, and it is also possible to discover new positives within the dataset. Based on our hypothesis, we provide a novel design of two popular contrastive loss functions, and explore their effectiveness on four heterogeneous state-of-the-art approaches. The extensive experimental analysis conducted on four datasets, EPIC-Kitchens-100, MSR-VTT, MSVD, and Charades, validates the effectiveness of the proposed strategy, observing, e.g., more than +20% nDCG on EPIC-Kitchens-100. Furthermore, these results are corroborated with qualitative evidence both supporting our hypothesis and explaining why the proposed strategy effectively overcomes it.

1. Introduction

When looking for a video via a multimedia search engine, the user usually describes its expected contents by means of a natural language query. Then, the multimedia search engine responds with a ranking list of visual items, which, according to the underlying decision-making system, best fit the contents described by the given user query. Notably, the output list may contain a multitude of visual items which are equally valid, i.e., highly relevant, to the input. Take for instance Fig. 1.a: given the query "two people are wrestling", the videos surrounded by the red solid line and the blue dashed line are both relevant to it. Previous works on text–video retrieval (e.g., Luo et al. (2022), Shvetsova et al. (2022) build a system retrieving the *only* video paired to a given query in the dataset. This setting is referred to as Instance-based Video Retrieval, or IVR. However, IVR neglects that other videos in the dataset, such as the one on the right in Fig. 1.a, could be relevant to the query, and their rank is as important as the groundtruth one. This limitation can lead to incomplete and unsatisfactory results for the user, as relevant information could be missed. A recent take on this problem is provided by Wray et al. (2021), who raised the awareness on the limitations of IVR and introduced the Semantic Similarity Video Retrieval (SSVR): by aiming at retrieving both the groundtruth caption and all the semantically equivalent captions, it ensures a comprehensive and accurate retrieval of all the relevant information.

Currently, the models for SSVR are built upon the foundation of IVR models, which are created by training on large datasets composed of video-caption pairs. Most of the state-of-the-art methods build a joint textual–visual embedding space via deep learning, e.g., Dong et al. (2021), Shvetsova et al. (2022). The underlying neural network learns to produce similar representations for a video clip and its associated textual description, thus aligning them in a joint visual–textual embedding space. This allows for the use of a textual query, mapped into the same embedding space, to obtain a ranking of all the videos, and vice versa the ranking of the captions can be obtained by using

[*] Corresponding author at: University of Udine, Via delle Scienze, 206, Udine 33100, Italy.
E-mail address: falcon.alex@spes.uniud.it (A. Falcon).

https://doi.org/10.1016/j.cviu.2024.104035
Received 28 November 2023; Received in revised form 8 May 2024; Accepted 14 May 2024
Available online 20 May 2024
1077-3142/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

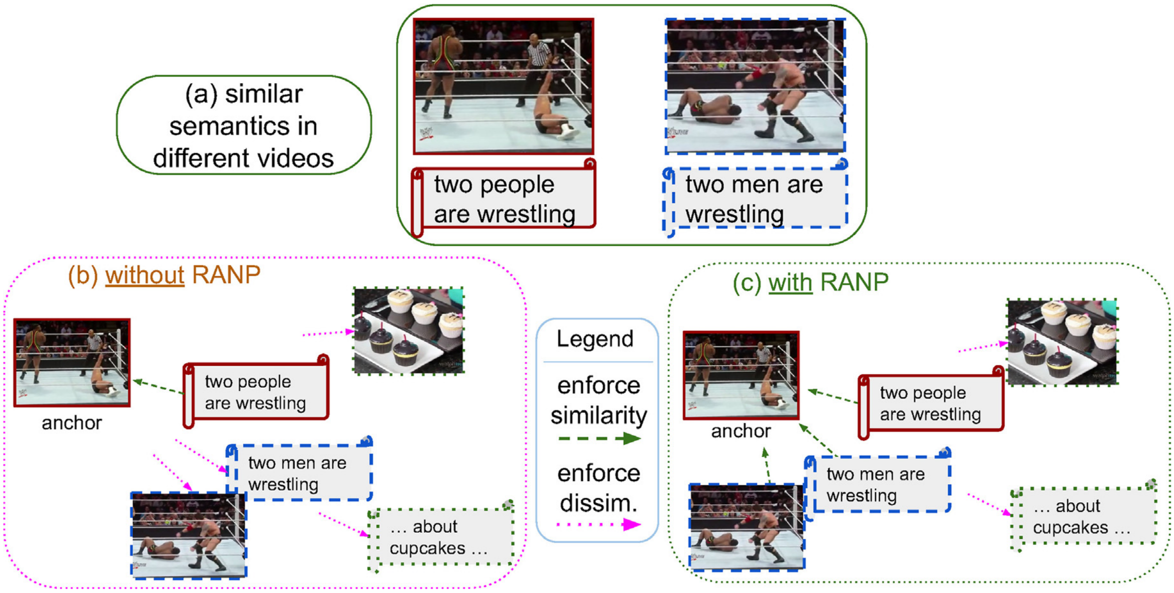


Fig. 1. Current cross-modal retrieval models are trained by enforcing similarity constraints on cross-modal elements only when they are paired together in the dataset, e.g., the *anchor* video surrounded by the red solid line and its groundtruth caption "two people are wrestling". (**a**) Large datasets often contain videos which naturally share *similar semantics*, e.g. both the two videos (and the respective captions) describe two professional wrestlers performing on the stage. (**b**) With current methodologies, dissimilarity constraints are enforced between the anchor and *all* the other videos and captions. However, the anchor video and its caption naturally share *similar semantics* with other samples, e.g., those surrounded by a blue dashed line, making it contradictory to force dissimilarity in their representations. (**c**) The proposed strategy overcomes this shortcoming, enforcing dissimilarity constraints with irrelevant samples (e.g., green dotted border), and similarity constraints with relevant ones (e.g., blue dashed border). The example videos and annotations are taken from MSR-VTT (Xu et al., 2016). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a video. To learn a model capable of producing representations in a joint text–video embedding space, a peculiar type of loss functions, called *contrastive* loss functions, is often used, e.g., Chopra et al. (2005), Oord et al. (2018) and Schroff et al. (2015). These functions aim at maximizing the similarity of videos and captions which are paired in the dataset, while minimizing that of unpaired examples, as shown in Fig. 1.b. Specifically, given a video, its own caption represents a *positive* example, whereas all the other captions serve as *negative* examples.

A great effort was spent by the community on developing methodologies to select the negative examples, e.g. by selecting — or *mining* — one (Schroff et al., 2015), two (Chen et al., 2017), or more negatives (Sohn, 2016). Other researchers focused on identifying which negatives contribute to the loss: in particular, those which contribute the most are called *hard* negatives, whereas those which contribute highly but are not more similar to the query than a positive example form the *semi-hard* negatives (Schroff et al., 2015). Notably, these approaches effectively train models extracting similar latent representations for paired inputs, making their utilization advantageous for obtaining state-of-the-art IVR performance. At the same time, they assume that unpaired videos and captions are never relevant to each other. However, this assumption hardly holds in real world scenarios: Fig. 1.a captures this situation, which is commonly found in many public datasets. By following such an assumption, a semantically similar caption (e.g., "two men are wrestling") could be chosen as a negative for the anchor video, therefore forcing the model to extract different representations for them (Fig. 1.b). Yet, by looking at the videos and by reading their captions, it is clear that no real difference in their semantics is present and they should not be contrasted.

Meanwhile, the community investigated less on the usage of positive examples mainly because of how video–text datasets are usually built: in fact, there is no groundtruth label to define two videos as semantically similar, as for those in Fig. 1.a. A first attempt, involving the creation of action labels, was proposed by Wray et al. (2019, 2021) to perform the mining of both positives and negatives *offline*, that is by selecting them without taking into account their contribution to the loss.

We hypothesize that these two issues represent major disadvantages of IVR models, limiting their understanding of semantics and their effectiveness in SSVR: given a video–caption pair, the negative samples for it may wrongly be videos or captions with highly similar contents that can harm contrastive learning, while the positives are not sampled and therefore under-represented during training although truly relevant videos and captions may be found within the available data. Therefore, in this paper we propose a novel strategy composed of two main components, effectively improving the *online* selection of both positives and negatives by leveraging the overlap of semantic concepts shared by the captions attached to the videos. By doing so, a model trained with the proposed strategy can overcome both the shortcomings of IVR methodologies, as illustrated in Fig. 1.c. First, it only selects the negatives which are not semantically similar to the anchor. Second, it can identify the captions which are not related to a video in the dataset but share similar semantics with it, and use them as positive examples in a properly reformulated loss function.

The proposed strategy is flexible and can be applied to different methods and loss functions. Specifically, we provide a novel reformulation of two popular contrastive loss functions, taking into account the hypothesis we formulated. Then, we experimented their use in four heterogeneous state-of-the-art methods for IVR. Regarding the loss functions, we consider the Triplet loss, which maximizes the cosine similarity of a query and a positive example, while enforcing a margin to the similarity between the query and one negative at a time (Schroff et al., 2015); and the NCE loss defined by Miech et al. (2020) which maximizes the similarity of the positive pair while minimizing that of all the negatives within the batch. We validate our strategy on a method using hierarchical learning and graph reasoning (Chen et al., 2020b), on a method using Transformers as a multimodal fusion technique (Shvetsova et al., 2022), and on two CLIP-based methods (Li et al., 2023; Luo et al., 2022). We conduct an extensive experimental analysis on four public datasets, MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), Charades (Sigurdsson et al., 2016), and EPIC-Kitchens-100 (Damen et al., 2021a). In particular, we observe consistent and considerable improvements across the many cases under analysis, e.g., more than +20% nDCG on EPIC-Kitchens-100 and +4% on MSR-VTT, overcoming previous state-of-the-art approaches. Moreover, we provide qualitative evidence supporting our hypothesis, and explain *why* the proposed strategy is so effective at overcoming it. Finally, to support reproducibility, we publicly release the trained models and the code supporting both methods and loss functions at https://github.com/aranciokov/ranp.

A preliminary version of this work was published as Falcon et al. (2022b). In this manuscript, we extend the previous work following two main directions: improving and making the solution more general, and performing an in-depth analysis of the embedding space learned by the proposed strategy.

In particular, we propose a novel way to include in our strategy the usage of both hard negatives, as in the previous work, and semi-hard and easier negatives. This extension allowed us to obtain better performance on MSR-VTT, obtaining a relative improvement of +13.8%.

Moreover, in this paper, we also generalize the formulation of our strategy to make it work with other loss functions and other methods. In particular, we formulated a novel definition of the NCE loss following the general idea proposed in our strategy and integrated the reformulated loss functions within three new state-of-the-art approaches based on Transformers and CLIP. In all the cases under analysis, using the proposed strategy at training time leads to improved performance compared to the use of the original loss function. For instance, on Everything-at-once, we observed a +20% and +7% increase in nDCG on EPIC-Kitchens-100 and MSR-VTT, respectively, and on CLIP4Clip, we saw a +1% increase in nDCG on MSR-VTT.

Finally, we perform an extensive qualitative analysis (Section 5.3) allowing us to explain *why* we observe significant improvements (e.g., more than +20% nDCG on EPIC-Kitchens-100) and why IVR methods do not perform well on SSVR.

The main contributions can be summarized as follows:

- We highlight two important limitations of IVR methods limiting their effectiveness on the SSVR task. First, given a video, IVR methods consider the groundtruth caption as the only relevant caption for that video, while videos or captions with highly similar contents are treated as irrelevant, potentially harming contrastive learning. Second, other relevant examples are not sampled and therefore under-represented during training although truly relevant videos and captions may be found within the available data. We hypothesize that these two limitations highly affect the final performance obtained on the SSVR task.
- We propose a general training strategy, RANP, to overcome these two shortcomings. RANP uses the overlap of semantic concepts shared by the captions to effectively improve the online selection of both positives and negatives used in contrastive learning. Then, we implement this general strategy by reformulating two popular training loss functions, obtaining the novel Triplet-RANP and NCE-RANP loss functions.
- To test our hypotheses, we integrate our loss functions into four highly heterogeneous approaches (HGR, Everything-at-once, CLIP4Clip, and ProST) and evaluate them on four public datasets (EPIC-Kitchens-100, MSR-VTT, MSVD, and Charades). The experiments include a vast range of quantitative and qualitative analyses. The results confirm the effectiveness of the strategy (e.g. +20% nDCG on EPIC-Kitchens-100 and +4% on MSR-VTT) and its robustness (e.g. consistent improvements across datasets, methods, and optimization strategies), while also showing why the proposed strategy is effective for the SSVR task.

After this introduction, in Section 2 we present the related works on cross-modal video retrieval and to contrastive losses. In Section 3, we describe how IVR methods are typically obtained, providing details about the important shortcomings which limit their semantics awareness and their effectiveness in SSVR. Then, to address these limitations, in Section 4 we introduce our training methodology, whereas several experiments and analyses are discussed in Section 5. Finally, Section 6 concludes the manuscript.

2. Related work

We start this section by analyzing the works related to Instance-based Video Retrieval (IVR), as they also serve as foundational components for Semantic Similarity Video Retrieval (SSVR). Then, we move to SSVR and analyze the main works on this topic. Finally, we shift the focus to contrastive losses, a fundamental topic for the methodology proposed in this paper.

Instance-based Text–Video Retrieval. Considering the recent introduction of SSVR, the methods presented for it rely on the foundations built for IVR, which has been extensively studied within the community. Currently, state-of-the-art IVR methods build a joint textual–visual embedding space in which retrieval is done by mapping the query and then by ranking the other candidates following a similarity metric, e.g., Shvetsova et al. (2022), Luo et al. (2022).

Since videos are composed of many modalities, many techniques to learn joint representations were introduced, e.g. Liu et al. (2019), Miech et al. (2018), Mithun et al. (2018), Gabeur et al. (2020), Wang et al. (2021) and Yang et al. (2022). For instance, MoEE (Miech et al., 2018) and T2Vlad (Wang et al., 2021) are based on NetVLAD (Arandjelovic et al., 2016), whereas Collaborative Experts (CE) (Liu et al., 2019) introduced a gating mechanism for the visual and audio-related features directed by several pretrained experts. Similarly, Teach Text (Croitoru et al., 2021) leveraged the availability of multiple language models to obtain multifaceted representations of the captions associated to the videos. With the advent of Transformers, they were more frequently employed to coordinate the experts and learn effective multimodal models, e.g., MMT (Gabeur et al., 2020) and Everything-at-once (Shvetsova et al., 2022).

Although multiple experts help understanding better the multiple types of information relevant for the video, structural inductive biases are used to impose preferences over the space of solutions, potentially leading to better generalization. To this end, several works focused on learning structured embeddings following the structure of the input data, e.g. by working on the part-of-speech (JPoSE, Wray et al. (2019)), by learning global and local representations via semantic roles (HGR, Chen et al. (2020b)), by describing complex queries via latent semantic trees whose nodes represent single words and constituents (Yang et al., 2020), or by learning a hierarchical representation on short- and long-term videos/paragraphs and clips/captions (Ashutosh et al., 2023).

More recently, driven by the availability of larger multimodal datasets and the scalability of Transformers, a different trend emerged, that is joint vision-and-language pretraining, simplifying the complexity of network architecture in favor of much larger pretraining. This led to important advancements in cross-modal understanding (CLIP, Radford et al. (2021)), which were also brought to the video community (Portillo-Quintero et al., 2021; Gao et al., 2021; Fang et al., 2021; Luo et al., 2022). In particular, Luo et al. (2022) presented CLIP4Clip and showed that the knowledge learnt with the CLIP objective is also greatly useful to achieve better video understanding capabilities. Additional pretext tasks were introduced to improve video-language pretraining, including the VideoQA-inspired Multiple Choice Questions by Ge et al. (2022), which requires the model to correctly answer noun and verb entities in masked questions, or the Prompting Entity Modeling (Li et al., 2022), designed to improve fine-grained alignment between visual regions of the frames and text entities found within the captions. ProST (Li et al., 2023) introduced a progressive local-to-global spatio-temporal modeling to avoid the loss of fine-grained spatial details while modeling the inter-frame dependency. Recently, significant advancements were obtained by adapting large language models to interact with multiple modalities, with the aim of obtaining larger scale pretraining video–text datasets using large language models to automatically generate and rephrase captions (Zhao et al., 2023), or generating new captions using large image-language models adapted for video tasks (Zhao et al., 2024).

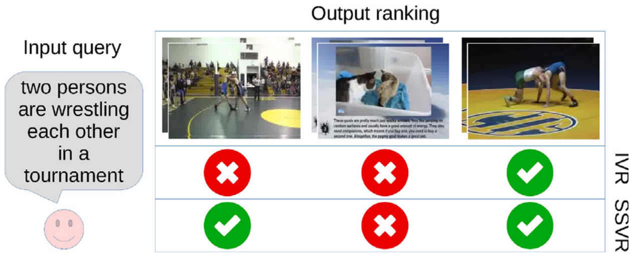


Fig. 2. Instance-based Video Retrieval (IVR) compared to the recently introduced Semantic Similarity Video Retrieval (SSVR). In IVR, the only video which needs to be retrieved correctly, i.e., at the top of the ranking list, is the one described by the input query in the dataset. In SSVR, multiple videos are relevant to the query and they all need to be retrieved following a descendant order of relevance, making this setting closer to real-world applications.

Currently, state-of-the-art IVR methods obtain high recall rates, making it likely to retrieve the groundtruth among the first ranks for any query (e.g., Luo et al. (2022)). However, their training procedure still suffers from a fundamental limitation: given a query, they only consider the groundtruth video as relevant, and, consequently, all the others are entirely irrelevant, overlooking the potential of instructing the model to retrieve other pertinent videos. This limitation highlights the need for a more comprehensive approach that aims at retrieving all semantically equivalent captions or videos.

Semantic Similarity for Video Retrieval. Recently, multiple works on cross-modal retrieval have highlighted the shortcomings of IVR which, given a caption (respectively, a video clip), only seeks to retrieve the groundtruth video (resp., caption), e.g., Chun et al. (2021), Wang et al. (2022), Wray et al. (2021) and Wu et al. (2022). In contrast, Wray et al. (2021) introduced the Semantic Similarity Video Retrieval problem (SSVR), which aims to retrieve *all* semantically equivalent videos (or captions). This setting is also closer to real-world applications, which require a deep understanding of semantics to retrieve all relevant videos and fulfill the user search. For instance, Fig. 2 presents a scenario in which the video paired to the input query is the 3rd video in the ranking list (on the right). In IVR, the top ranked video (on the left) is a false positive, since it is described by a slightly different caption. However, in IVR it does not count towards a successful retrieval, although the user could still be satisfied with its content (both videos depict a wrestling match). Differently from IVR, the top ranked video is a good match for SSVR, as its content is relevant to the query despite the mismatched caption.

As initial efforts to address these limitations, Chun et al. (2021) leveraged probabilistic distributions to capture uncertainty over the one-to-many correspondences, whereas Wang et al. (2022) achieved similar goals by capturing richer semantics via the projection of the images and captions into rectangular areas of the embedding space which contain semantically related elements. Falcon et al. (2022c) adapted the triplet loss function by reformulating the fixed margin in terms of the relevance. Differently from them, in our work, we aim to bridge the gap between IVR and SSVR by means of semantic relations between different video and caption pairs in the dataset which we automatically discover.

Contrastive loss and mining techniques. Contrastive losses (Gutmann and Hyvärinen, 2010; Hadsell et al., 2006; Hermans et al., 2017) are often used for cross-modal tasks because of their capability to maximize the descriptors' similarity for video and caption pairs in the dataset. Early works computed the loss on two samples at a time (Hadsell et al., 2006), whereas triplets (Schroff et al., 2015), quadruplets (Chen et al., 2017), and 'N+1'-tuples (Sohn, 2016) were later used. Yet, training on all the possible tuples from the dataset is unfeasible (e.g. the amount scales cubically with triplets), and many of them may not even contribute to the loss. Therefore, a subset of the tuples are selected through mining techniques, either from the dataset ('offline') or from the batch ('online').

The former is often avoided because of the need to re-sample them during training, making it burdensome. Nonetheless, it was used in several domains, e.g., deep metric learning (Harwood et al., 2017; Suh et al., 2019) and video retrieval (Wray et al., 2019, 2021).

Differently, online mining forms the tuples within the batch and is widely used (Croitoru et al., 2021; Dong et al., 2021; Shvetsova et al., 2022). In the literature on video retrieval, the negatives for a video are simply all the other videos and captions which are not paired to it. Then, the loss is usually computed either on all negatives, e.g., Gabeur et al. (2020), Miech et al. (2018), despite it leading to extra computation, or on a subset of them, such as those which share a highly similar representation to the positive pair (Chen et al., 2020b; Dong et al., 2021). Nonetheless, recent research also presented the usefulness of easy examples (Xuan et al., 2020a,b). On the other hand, positive examples are not mined for video retrieval, and the only positive caption for a video is its own. In some fields, e.g., in cross-modal (Hermans et al., 2017; Xuan et al., 2020b) and near-duplicate video retrieval (Jiang et al., 2019), additional positive examples were also mined, yet they use labels available in the dataset. For instance, SVD (Jiang et al., 2019) contains 1206 videos in the query set for which more than 10000 video pairs are labeled as positive. An attempt which creates action labels, was proposed for offline mining in Wray et al. (2021). In other research fields, such as in representation learning for images or videos, positive samples were also created via transformations (Chen et al., 2020a; He et al., 2020; Pan et al., 2021; Qian et al., 2021).

Differently from previous works, we introduce semantic knowledge to the training process by computing an overlap of the semantic concepts shared among videos and captions. Moreover, we devise a two-step method to discover new positives within the batch and use them to improve the training.

3. Training an instance-based video retrieval model with contrastive loss and mining

Given a video $v^\star$ and a pool of candidate textual descriptions, the objective of video-to-text retrieval is to orderly retrieve each of the descriptions based on how well they describe the video, thus producing a ranking list in which such a order is given by a similarity function (computed with $s(\cdot, \cdot)$, e.g., cosine similarity). IVR implements this problem in a instance-based way, prioritizing the retrieval of the ground truth caption, $q^\star$, as the top ranked candidate, while neglecting that multiple captions may be equally valid for a given video. Unfortunately, this means that the quality of the ranking list is not taken into consideration, possibly leading to unsatisfactory results (see Fig. 2).

To implement an IVR system, text–video retrieval models are often trained by means of a contrastive loss, aiming at the maximization of similarity of the descriptors computed for pairs of visual and textual data. To do so, the triplet loss (Schroff et al., 2015) and the NCE loss (as implemented by Miech et al. (2020)) are common choices (e.g., Chen et al. (2020b), Luo et al. (2022) and Shvetsova et al. (2022)). Specifically, the NCE loss achieves this goal by averaging the cost computed for each video–text pair in the batch. The cost for a single pair is given by:

$$L_{NCE}(v^\star, q^\star) = -log \frac{\exp(v^{\star T} q^\star / t)}{\sum_{i=1}^{B} \exp(v^{\star T} q_i / t)} \tag{1}$$

where $v^\star$ and $q^\star$ are the paired data, t is a temperature parameter, and the softmax normalization is computed with respect to the B captions in the batch. An analogous term, $L_{NCE}(q^\star, v^\star)$, is computed when ranking videos instead of captions for text-to-video retrieval. Differently from the NCE loss, the triplet loss computes the cost for a video–text pair by also considering the negative elements explicitly. The cost given by the video–text pair and one of the negatives is computed as follows:

$$L_n(v^\star, q^\star, q-) = max(0, \Delta_n + s(v^\star, q-) - s(v^\star, q^\star)) \tag{2}$$

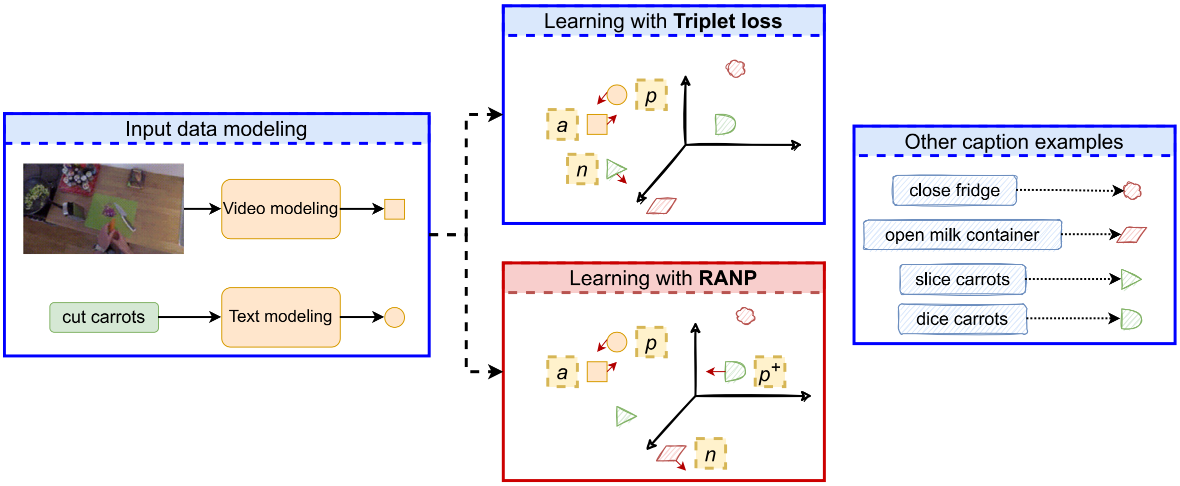


Fig. 3. By adopting the typical approach, a caption which is not paired to the input video is selected as hard negative based solely on its similarity. Yet, this may lead to semantically similar captions being wrongly selected as negatives, despite their high relevance to the video (see Section 3). With our proposed technique, RANP (see Section 4), we avoid this while also finding new positive captions and, consequently, the learning also increases their similarity to the video.

where Δ_n is a fixed margin, and $q-$ represents the negative caption. By optimizing Eq. (2), Δ_n is enforced between $s(v^\star, q^\star)$ and the similarity between the video and the negative query, in order to satisfy the following constraint:

$$s(v^\star, q-) + \Delta_n < s(v^\star, q^\star) \quad (3)$$

The optimization of Eq. (2) can be performed in several ways. Typically, all the negative captions in the mini-batch are used, i.e., Eq. (2) is accumulated for all the negatives $q-$ and the video–text pairs, then the average is taken. For instance, it is done in Gabeur et al. (2020), Miech et al. (2018). However, this means that the loss for many easy negative captions, i.e. already satisfying Eq. (3), is computed although they do not contribute meaningfully to it. Therefore, the selection of a single negative example per video–text pair is often preferred. To this end, the online hard negative mining selects the most similar example to the anchor within the batch, e.g., in Chen et al. (2020b), Dong et al. (2021). While these examples are informative to the training process, their usage from the very start may lead the optimization process to a local minimum where the model collapses (Schroff et al., 2015). To avoid it, *semi-hard* negatives, i.e. highly similar to the anchor but less than the positive example, are often preferred and can be used also to start the training process (Hermans et al., 2017; Schroff et al., 2015).

However, all these sampling techniques and both the NCE loss and the triplet loss, which work greatly for IVR, are affected by the same limitations highlighted in Section 1: they always treat unpaired videos/captions as irrelevant to each other even though they share similar semantics, and they neglect the existence of additional positives examples. We hypothesize that these two limitations strongly limit the semantics understanding and the SSVR performance of state-of-the-art IVR methods. The two limitations are formalized in Sections 3.1 and 3.2. After that, Section 4 introduces the proposed method.

3.1. Limitation: relevant videos and captions are treated as completely irrelevant negatives

As mentioned in the previous section, both Eq. (1) and (2) consider all q_i different from $q^\star$ as negatives examples for a given $v^\star$. This means that, if Q identifies the set of captions in the mini-batch, then the set $\mathcal{N}_{v^\star} = Q \setminus \{q^\star\}$ captures the negative captions for $v^\star$. Yet, $\mathcal{N}_{v^\star}$ may still contain captions which should not be considered negative because they describe in part or entirely the content of the video clip. For instance, let $q^\star$ be 'cut carrots', q_1 'slice carrots', q_2 'open milk container', and $s(v^\star, q_1) > s(v^\star, q_2)$ as in Fig. 3. Since it is not $q^\star$, q_1 could be chosen as a negative. Intuitively, the IVR assumption forces the model to discriminate 'cut carrots' and 'slice carrots' at training time, although those two actions are visually the same. Meanwhile, it implicitly considers both 'open milk container' and 'slice carrots' as equally irrelevant to 'cut carrots'. However, in SSVR q_1 and q_2 should be regarded as having different relevance values to $v^\star$, and q_1 should be excluded from the pool of candidate negatives. With our techniques, these bad selections are avoided from the start by using the semantics of the data. Conversely, existing methods only exclude $q^\star$ and let $s(\cdot,\cdot)$ guide the mining, making these situations likely to happen all the time during training.

3.2. Limitation: positive videos and captions are not sampled

As detailed in Section 2, when training a text–video retrieval model the positive examples are not mined, due to how text–video datasets are created, i.e., by having as the only 'labels' the association between a video and its own descriptions. Therefore, the model is unaware that additional valid video–caption pairs may exist within the dataset. This is a major shortcoming in the context of the SSVR task. As mentioned before, $\mathcal{N}_{v^\star}$ may contain captions suitable to become new descriptions for the video, thereby providing semantic supervision. Notably, these additional captions can be interpreted as new annotations obtained through semantic-preserving transformations. For instance, 'slice carrots' (q_1 in the example made in Section 3.1) may be obtained from 'cut carrots', originally describing $v^\star$, making it a valid positive which could be used at training time to learn useful information for the SSVR task. With the proposed methodology, we aim at identifying these situations and use them at training time, without needing any additional labeling.

4. Relevance-aware online mining of positives and negatives

To improve the selection of the negatives and the positives for the SSVR task, we leverage the shared semantics of videos and captions. In particular, our approach is based on the quantification of the overlap of shared semantic concepts, which are identified in terms of nouns, verbs, and their synonyms. As an example, let: (x_1) 'pick up a flowerpot and a sunflower', (x_2) 'pick an helianthus and a flowerpot', (x_3) 'pot the lily in a flowerpot', (x_4) 'put the cake in the oven'. Since 'helianthus' and 'sunflower' are synonyms, x_2 and x_1 are semantically the same (hence, x_2 is highly relevant to x_1); x_3 is slightly relevant because of 'flowerpot', but the flowers and actions are different; and x_4 is irrelevant. Thus, we seek a notion of relevance which captures semantic relations, such as synonyms, and employ it to determine a continuous value representing the semantic closeness between two captions. For such a task, we adopt

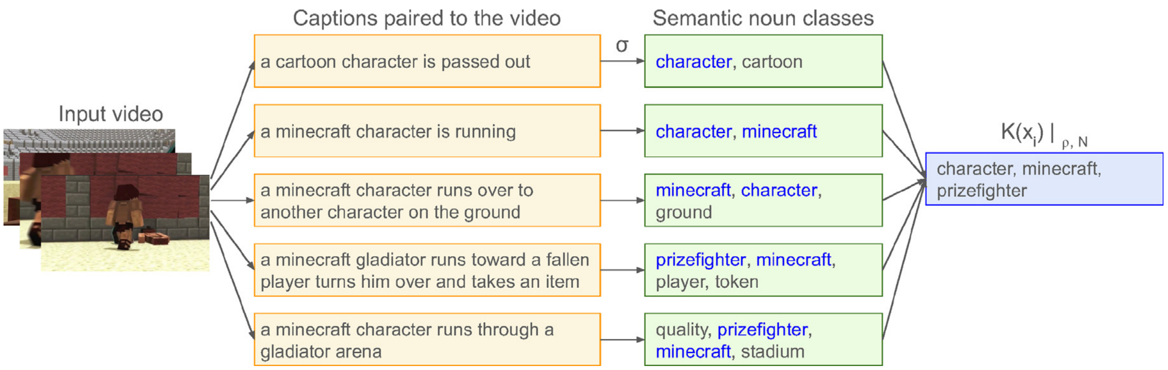

Captions paired to the video
Semantic noun classes
Input video
σ
a cartoon character is passed out
a minecraft character is running
a minecraft character runs over to another character on the ground
a minecraft gladiator runs toward a fallen player turns him over and takes an item
a minecraft character runs through a gladiator arena
character, cartoon
character, minecraft
minecraft, character, ground
prizefighter, minecraft, player, token
quality, prizefighter, minecraft, stadium
K(x_i) | ρ, N
character, minecraft, prizefighter

Fig. 4. Construction of the noun word set $\mathcal{K}(x_i)_{|\rho,N}$ for an example video x_i with five captions (as an example, $\rho = 0.40$). Classes which are selected for the word set are highlighted in blue. Details in Section 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the definition for a relevance function $\mathcal{R}(x_i, x_j)$ given by Damen et al. (2021a):

$$\mathcal{R}(x_i, x_j) = \frac{1}{2}\left(\frac{|x_i^V \cap x_j^V|}{|x_i^V \cup x_j^V|} + \frac{|x_i^N \cap x_j^N|}{|x_i^N \cup x_j^N|}\right) \tag{4}$$

where x_i^V and x_i^N represent, respectively, the set of verb and noun classes identified in the ith caption. Classes are used here to include both a token, e.g., 'helianthus', and its synonyms, e.g., 'sunflower'. When x_i (or x_j) is a video, two situations arise, based on the size of $\mathcal{D}(x_i)$, i.e., the captions paired to the video. If it is paired to only one caption q_i in the dataset, then the noun and verb classes of q_i are also used for the video. Conversely, if there are multiple captions, then a word set is built by picking the classes which are shared among many captions of the video itself, as in Wray et al. (2021). Formally, given x_i, its word set for nouns is defined as $x_i^N = \mathcal{K}(x_i)_{|\rho,N}$, representing the set of semantic noun classes appearing in at least $\rho \cdot |\mathcal{D}(x_i)|$ captions. By doing so, we intuitively annotate a video with the nouns mentioned by multiple captions, hence leading to a more robust word set. Formally, $\mathcal{K}(x_i)_{|\rho,N}$ is defined as $\{c \mid \sigma(c) = N \land |\{d | d \in \mathcal{D}(x_i) \land c \in d\}| \geq \rho \cdot |\mathcal{D}(x_i)|\}$, that is, it comprises the classes c which represent nouns ($\sigma(c) = N$) and the amount of captions d containing the class c ($\{d | d \in \mathcal{D}(x_i) \land c \in d\}$) is at least ρ% of the total captions of the video ($|\{d | d \in \mathcal{D}(x_i) \land c \in d\}| \geq \rho \cdot |\mathcal{D}(x_i)|$). Fig. 4 illustrates these steps. The same steps are used for x_i^V. Finally, looking at the example, the following are computed: $\mathcal{R}(x_1, x_2) = 1$, $\mathcal{R}(x_1, x_3) = 0.16$, and $\mathcal{R}(x_1, x_4) = 0$.

In the following Sections, we introduce the two main components of our sampling strategy. The first one (Section 4.1) solves the shortcoming related to relevant videos and captions being treated as completely irrelevant candidates. The second component (Section 4.2) introduces a novel strategy to sample relevant positive candidates and uses them at training time, addressing the lack of positive mining in text–video retrieval approaches. Notably, as the proposed strategy acts at training time, it does not impact on the complexity (in terms of FLOPs, parameters count, or inference time) of the underlying methods. The major burden is introduced during the preprocessing, since the captions need to be parsed and analyzed to obtain part-of-speech tags and then the semantic classes. At training time, there is an increase in time required for processing the batch, especially for computing the loss. For instance, using HGR on EPIC-Kitchens-100, we observed an average of 34.3 ms to process one batch (n = 1050 batches), whereas the average time is 13.4 ms for the baseline (n = 1050).

4.1. Relevance-aware online mining of negative examples

In Section 3.1, we provide an intuitive description of a shortcoming of current techniques used to mine negative examples. Formally, if we define a threshold τ to separate irrelevant from relevant content, then the relevant captions, i.e. $\{q \mid \mathcal{R}(v^\star, q) \geq \tau\}$, may have a non-empty intersection with the negative captions $\mathcal{N}_{v^\star}$: as a consequence, a relevant caption may be selected as a negative, leading to the shortcoming mentioned in Section 3.1. Note that the same problem affects both the Triplet loss and the NCE loss. We address this issue by introducing RAN, which binds the sampling procedure to the relevance function, hence avoiding the selection of a 'false negative'. This is done by removing the relevant content from the negatives' pool, obtaining a set of truly irrelevant captions as follows:

$$\mathcal{N}_{v^\star}^{\mathcal{R}} = \mathcal{N}_{v^\star} \setminus \{q \mid \mathcal{R}(v^\star, q) \geq \tau\} \tag{5}$$

which we treat as the 'negative' candidates pool. Therefore, the relevance function $\mathcal{R}$ becomes fundamental in the sampling procedure, and the exclusion of an example is no longer based solely on its relation to $v^\star$ in the dataset.

4.2. Relevance-aware online hard positive mining

By using RAN, only irrelevant samples are used as negatives, and relevant ones are not seen as negatives anymore. Yet, relevant captions and videos could still play a role in the optimization process but, as mentioned in Section 3.2, they are not currently used. Therefore, we propose RANP, a two-step strategy to discover additional relevant samples, thus adding positive mining which is not pursued for text–video retrieval. To do so, the first step consists in finding a relevant caption $q+$ for $v^\star$, *i.e.* $\mathcal{R}(v^\star, q+) \geq \tau$, which has a far too dissimilar representation when compared to $v^\star$. Notably, we look for such a $q+$ because the model is unaware of the semantic content it shares with $q^\star$ and $v^\star$, since it creates a dissimilar representation for $q+$: hence, it requires the additional supervision we aim to provide via this strategy. To mine $q+$, a 'pool of positives' $\mathcal{P}$ needs to be established, since $Q \setminus \{q^\star\}$, i.e., $\mathcal{N}_{v^\star}$ may contain irrelevant captions. By using $\mathcal{R}$ to compute the relevance, we define $\mathcal{P}_{v^\star}^{\mathcal{R}}$ by *excluding* the irrelevant samples found in $\mathcal{N}_{v^\star}^{\mathcal{R}}$: $\mathcal{P}_{v^\star}^{\mathcal{R}} = Q \setminus \mathcal{N}_{v^\star}^{\mathcal{R}}$. Then, $q+$ is sampled as follows:

$$q+ = argmin_{q \in \mathcal{P}_{v^\star}^{\mathcal{R}}}\ s(v^\star, q) \tag{6}$$

Since $Q \setminus \mathcal{N}_{v^\star}^{\mathcal{R}} = \{q \mid \mathcal{R}(v^\star, q) \geq \tau\}$, the positives' pool will contain only those samples whose relevance with $v^\star$ is greater than τ.

The second step involves the addition of a new term to the loss function, which aims at increasing the similarity of $v^\star$ to the newly discovered $q+$. Since this step is related to the loss function under analysis, we reserve two subsections (4.2.1 and 4.2.2) to the two popular loss functions we considered in this study, the Triplet loss and the NCE loss, although it can be seen that the extension of the proposed idea to additional functions is straightforward.

4.2.1. Triplet-RANP

For a given $v^\star$, the Triplet loss samples one or more negative captions from $\mathcal{N}_{v^\star}$ (see Section 3.1) and then computes the loss by using Eq. (2). To integrate RANP into the Triplet loss, we first apply RAN, that is the sampling of the negatives is done on $\mathcal{N}_{v^\star}^{\mathcal{R}}$. For instance, this means that the hardest negative caption is identified by the following equation:

$$q- = argmax_{q \in \mathcal{N}^{\mathcal{R}}_{v^\star}} s(v^\star, q) \tag{7}$$

To integrate the novel positive mining provided by RANP, the loss is extended by including a new term which aims at increasing the similarity of $v^\star$ and $q+$. This can be done by introducing an additional triplet loss, where $q+$ plays the role of $q^\star$. Formally, the cost for a single triplet composed of video $v^\star$, positive caption $q+$, and negative caption $q-$ is computed as:

$$L_p(v^\star, q+, q-) = max(0, \Delta_p + s(v^\star, q-) - s(v^\star, q+)) \tag{8}$$

Finally, given a mini-batch of B clip and caption pairs, the final Triplet-RANP loss is computed by summing the video-to-text $\mathcal{L}_{v-t}$ and the text-to-video loss $\mathcal{L}_{t-v}$. In particular, $\mathcal{L}_{v-t}$ is defined as:

$$\mathcal{L}_{v-t} = \frac{1}{|B|}\Big(\sum_{\substack{v^\star \in B \\ q+ \sim \text{Eq. (6)} \\ q- \sim \text{Eq. (7)}}} L_p(v^\star, q+, q-) \; + \sum_{\substack{v^\star \in B \\ q- \sim \text{Eq. (7)}}} L_n(v^\star, q^\star, q-)\Big) \tag{9}$$

whereas $\mathcal{L}_{t-v}$ is computed by switching v and q.

4.2.2. NCE-RANP

Since the NCE loss does not have an explicit step used to identify the negative examples (see Eq. (1)), the integration of RANP is focused on the additional positive mining aspect. As mentioned in Section 3, Eq. (1) maximizes the similarity of $q^\star$ to $v^\star$, while it minimizes the similarity with all the other q_i. To integrate RANP into the NCE loss, we reformulate how NCE computes the cost for a video–text pair as follows, aiming to increase the similarity of $q+$ (Eq. (6)):

$$L_{NCE,p}(v^\star, q+) = -log \frac{\exp(v^{\star T} \mathrm{q}+/t)}{\sum_{i=1}^{B} \exp(v^{\star T} q_i/t)} \tag{10}$$

Then, we obtain the NCE-RANP video–text loss $\mathcal{L}_{NCE-RANP,v-t}$ by defining the following equation:

$$\mathcal{L}_{NCE-RANP,v-t} = \frac{1}{|B|} \sum_{\substack{v^\star \in B \\ q+ \sim \text{Eq. (6)}}} \big(L_{NCE}(v^\star, q^\star) + L_{NCE,p}(v^\star, q+)\big) \tag{11}$$

which is summed to $\mathcal{L}_{NCE-RANP,t-v}$ in order to obtain the final NCE-RANP loss.

5. Experimental results

We validate our methodology on four large scale vision and language datasets: EPIC-Kitchens-100, MSR-VTT, MSVD, and Charades. **EPIC-Kitchens-100** (Damen et al., 2021a) provides 67217 clips for training and 9668 for testing. A set of 4834 clips from the training set is used for validation and fast assessment of the performance, as done in the retrieval baselines of Damen et al. (2021a). Each clip is annotated by a brief caption describing an activity in the kitchen, and by verb and noun semantic classes. **MSR-VTT** (Xu et al., 2016) comprises 10000 clips about multiple scenarios, each annotated by 20 free-form captions. We follow the official split (from Xu et al. (2016)) of 6513, 497, and 2990 clips for training, validation, and testing. **MSVD** (Chen and Dolan, 2011) consists of 1870 open domain videos, each annotated with around 40 captions. The videos are split in 1200, 100, and 670 for training, validation, and testing, respectively. **Charades** (Sigurdsson et al., 2016) consists of 9848 videos, each annotated by multiple free-form descriptions, split in 7985 videos for training and 1863 for validation. For MSR-VTT, MSVD, and Charades, we compute the semantic classes with a pipeline σ made of spaCy, WordNet (Miller, 1995), and Lesk algorithm (Lesk, 1986) as in Wray et al. (2021). We use $\rho = 0.25$ (see Section 4).

We consider four recent methods for the experiments. **HGR** (Chen et al., 2020b) performs graph reasoning on hierarchical representations of video and caption. **Everything-at-once** (Shvetsova et al., 2022) is a recent Transformer-based method for text–video retrieval which uses multiple losses for aligning several modalities simultaneously. **CLIP4Clip** (Luo et al., 2022) performs pretraining with CLIP and then finetunes the underlying architectures (Transformer for text, Vision Transformer for visual data) for video understanding. **ProST** (Li et al., 2023) decouples the spatial and temporal modeling of video, and progressively aligns them in a local-to-global fashion with the textual features.

On both datasets, the training with HGR and Everything-at-once in its text–video version lasts for 50 epochs with a batch size of 64. For EPIC-Kitchens-100, we use officially provided TBN features (Damen et al., 2021a), whereas for MSR-VTT we use ImageNet-pretrained ResNet-152 features and Kinetics400-pretrained 3D-ResNeXt-101 features from Shvetsova et al. (2022). For MSVD and Charades, we use ImageNet-pretrained ViT-H-14 (Dosovitskiy et al., 2020) features extracted at 3fps using the standard library torchvision. We also extract Swin3D (Liu et al., 2022) features for Everything-at-once. In the case of CLIP4Clip and ProST, we follow the hyperparameters chosen by the authors on the ViT-B/32 backbone, and the training lasts for 5 epochs with a batch size of 128 (64 in the case of MSVD and Charades).

The evaluation on the testing set is performed with the best validation model. The metrics used for evaluation in the SSVR task are the Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002) and the Mean Average Precision (mAP) (Baeza-Yates et al., 1999), as in Wray et al. (2021). When not specified, nDCG is computed using the relevance function (Eq. (4)) with synset-aware semantic classes. In some experiments, it is also specified which semantic proxy (Wray et al., 2021) was used to compute the relevance: nDCG-SYN uses synset-aware classes; nDCG-BoW computes a single Intersection-over-Union (IoU) between sets of non-stop words in each caption; nDCG-PoS takes the average of two separate IoU for verbs and nouns lemmas (not semantic classes). For MSR-VTT, MSVD, and Charades we do not use mAP: because the computation of the semantic classes for videos consistently results in relevance values below one, even for ground truth annotations, the Average Precision is zero in the majority of cases. Consequently, the interpretation of the mAP lacks meaningful relevance from a retrieval perspective. For a comparison with IVR methods, we also use the recall rate (especially, R@1 and R@10) and the Geometric Mean of Recalls (GMR) used by Albanie et al. (2020).

The rest of the experimental section is organized as follows. Section 5.1 serves as a preliminary analysis to answer a fundamental research question: why is it important to consider semantic awareness when mining the negatives? Then, an in-depth quantitative analysis is done in Section 5.2, aiming to answer several research questions related to the analysis of the proposed strategy, including its cooperation with different optimization strategies and large scale pretraining, ablation studies, and a comparison with state-of-the-art solutions. After that, Section 5.3 analyzes the results from a qualitative point of view, explaining why the proposed strategy overcomes the shortcomings highlighted in Section 1. Finally, in Section 5.4, we discuss the limitations of our approach and possible future directions.

5.1. Why is it important to consider semantic awareness when mining the negatives?

In this experiment, we test the hypothesis that in the standard IVR methodology, relevant videos and captions are actually used as hard negatives. If confirmed, then it would mean that IVR methodologies force the model to create dissimilar representations even for samples which share the same semantic contents. We explore this question on the datasets under analysis. Fig. 5 presents the distribution of relevance values among the hard negatives selected in one epoch of training, hence containing multiple batches (e.g., more than 1000 in the case of EPIC-Kitchens-100, considering a batch size of 64) randomly constructed by the data loading process.

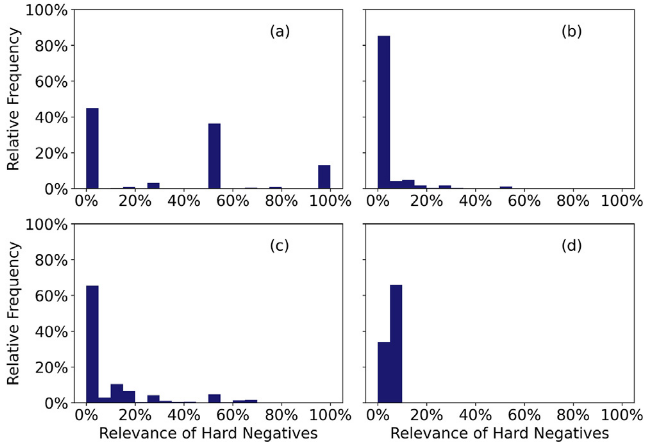


Fig. 5. Distribution of relevance values among the hard negatives mined in one epoch (batch size 64) on EPIC-Kitchens-100 (**a**), MSR-VTT (**b**), MSVD (**c**), and Charades (**d**).

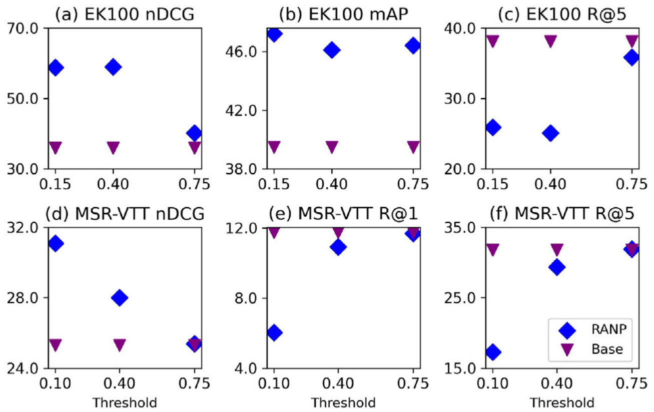


Fig. 6. We compare the influence of the threshold τ on the performance obtained by HGR. On EPIC-Kitchens-100: (a) nDCG (b) mAP (c) average R@5; on MSR-VTT (d) nDCG (e) average R@1 (f) average R@5. Discussion in Section 5.2.1.

On EPIC-Kitchens-100 (Fig. 5.a) more than 50% of the negatives have a positive relevance to the input caption, and 13% of them are 100% relevant, that is they share the same noun and verb classes. Overall, four modes for the relevance values are identified: 0 (with a relative frequency of 45%), 50 (36%), 100 (13%), 25 (3%).

As mentioned in Section 4, for MSR-VTT, MSVD, and Charades the video classes are chosen among those which appear in at least ρ, e.g. 25%, of the captions paired to that clip. Therefore, even the captions paired to the video may not have a relevance value of 100%: thus, finding relevant samples within random mini-batches is more difficult leading to much lower modes. In the case of MSR-VTT (Fig. 5.b), the modes are 0 (with a relative frequency of 85.2%), 10 (4.7%), 5 (4%), 15 (1.7%). For MSVD, 0 (65.4%), 10 (10.6%), 15 (6.5%), and 50 (4.8%). For Charades, the relevance values are always lower than 10%, highlighting only two modes: 0 (33.9%) and 5 (66.1%).

To conclude, it is confirmed that standard IVR methodologies consider relevant samples as entirely irrelevant.

5.2. Quantitative analysis

In this section, an in-depth quantitative analysis is performed. First, we present evidence of the impact on the performance of τ (Section 5.2.1), the training optimization strategy (Section 5.2.2), and large scale instance-based pretraining (Section 5.2.3). Then, Section 5.2.4 analyzes the loss behavior and the validation performance, highlighting how IVR losses lead to suboptimal results in SSVR. After conducting ablation studies in Section 5.2.5, we present a state-of-the-art comparison in Section 5.2.6.

5.2.1. How much does the threshold affect the final performance?

In Fig. 6, we report the results obtained by applying the Triplet-RANP with the three values for τ highlighted by the preliminary analysis (Section 5.1). Specifically, for EPIC-Kitchens-100 we visualize (6.a) the nDCG and (6.b) the mAP, and the average R@1 (6.c) whereas for MSR-VTT (6.d) the nDCG, and the average R@1 (6.e) and R@5 (6.f). On both datasets we keep $\Delta_n = \Delta_p = 0.2$ (as in Chen et al. (2020b)), since by changing Δ_n and Δ_p only small changes in performance are achieved.

According to the analysis in Section 5.1, lower values of τ should avoid the selection of several 'bad negatives' and help the training process. This is confirmed by Fig. 6, since the lower the value for τ, the higher the performance on both nDCG and mAP. In particular, on EPIC-Kitchens-100, it improves by up to +23.1% nDCG (with $\tau = 0.40$) and +7.7% mAP (with $\tau = 0.15$), whereas up to +5.8% nDCG is observed on MSR-VTT with $\tau = 0.10$. Notably, when very few examples are affected by the selection process, the performance obtained with the proposed strategy are similar to that obtained by the baseline (e.g., on MSR-VTT considering $\tau = 0.75$, see Fig. 6.d–f).

The situation is different if IVR metrics are considered. Two observations are made. First, if there are many relevant samples (that is, the threshold used at training time is small, e.g. $\tau = 0.10$) then it becomes more difficult for the learned model to identify the groundtruth among them and rank it accordingly, leading to lower recall rates even with large values of K. For instance, in EPIC-Kitchens-100 with HGR we see around 12% difference in R@5 (Fig. 6.c), which persists even with K = 100 (75.6% baseline, compared to 62.7%). Second, if there are fewer relevant samples (that is, the threshold is great, e.g. $\tau = 0.75$) then recall rates also become greater, even with small values of K. For instance, in EPIC-Kitchens-100, the difference in R@5 is much smaller (Fig. 6.c); in MSR-VTT, the difference is almost negligible both in R@5, obtaining 29.35% with $\tau = 0.40$ and 31.91% with $\tau = 0.75$, compared to 31.78% of the baseline (Fig. 6.f); and also in R@1, obtaining 10.92% ($\tau = 0.40$) and 11.69% ($\tau = 0.75$) compared to 11.75% of the baseline (Fig. 6.e).

Due to the focus on SSVR, in the following experiments we use $\tau = 0.15$ for EPIC-Kitchens-100 and $\tau = 0.10$ for MSR-VTT.

5.2.2. How much does the optimization strategy affect the final performance?

In Section 3, we mentioned that there are multiple approaches to mine the negatives and use either all or some of them to optimize the loss. To investigate how the selection of these strategies impact the performance and examine their interaction with the additional constraints introduced by RANP, we explore four of them: *All* refers to the usage of all the negatives; *Semi+All* means that semi-hard negatives are initially used, followed by an additional training with the All strategy; *Semi+Hard* starts with semi-hard negatives, and then uses hard negatives; finally, *Hard* means that hard negatives are used from the very start. Note that in our experiments with Triplet-RANP, the *Hard* strategy is only used with HGR, as it leads Everything-at-once to a collapsed model. This process differs from Curriculum Learning: in our strategy, we use all the examples in the training set during each iteration, whereas techniques based on Curriculum Learning (Soviany et al., 2022) do not, as the examples are partitioned in "difficulty classes" and selected based on the current model performance. Moreover, in Curriculum Learning the examples are considered difficult ("hard") by a per-task criterion, whereas "hard negatives" are difficult with respect to the loss function and a specific video–caption pair.

The results are presented in Table 1. Three major observations can be made. First, in all the considered cases the improvement obtained by RANP is consistent, across both SSVR metrics, models, and datasets. Second, on MSR-VTT the best results are achieved when using the

Table 1
Influence of the negative selection strategy in the triplet loss (discussed in Section 5.2.2). The symbol ✓is used to mark the usage of RANP during training.

Opt	Triplet-RANP	MSR-VTT		MSVD		EPIC-Kitchens-100			
Everything-at-once		nDCG	R@1	nDCG	R@1	nDCG-SYN	nDCG-BoW	nDCG-PoS	mAP
All		24.8	**6.9**	35.4	**31.5**	34.5	**27.7**	33.5	35.0
All	✓	**33.6**	4.2	**38.9**	24.3	**57.7**	27.3	**40.2**	**41.6**
Semi+All		21.3	**8.2**	35.1	**31.7**	33.3	26.7	32.9	33.9
Semi+All	✓	**34.4**	6.7	**41.2**	25.6	**58.6**	**31.2**	**43.8**	**46.0**
Semi+Hard		20.6	8.2	34.5	**31.9**	32.7	26.8	32.5	33.5
Semi+Hard	✓	**29.9**	**8.6**	**40.4**	24.2	**59.5**	**32.3**	**43.9**	**45.1**
HGR		nDCG	R@1	nDCG	R@1	SYN	BoW	PoS	mAP
All		26.7	**8.4**	37.2	**33.6**	37.1	**29.8**	34.6	40.8
All	✓	**34.4**	4.8	**41.4**	20.5	**57.5**	26.3	**41.1**	**42.4**
Semi+All		26.0	**9.9**	**37.0**	33.9	34.9	27.9	32.8	39.1
Semi+All	✓	**35.4**	7.5	36.4	**37.9**	**55.6**	**28.0**	**44.2**	**42.8**
Semi+Hard		23.8	**11.2**	41.5	26.6	34.4	27.2	32.5	38.1
Semi+Hard	✓	**27.8**	9.8	**43.2**	**32.2**	**54.0**	**31.3**	**43.0**	**45.5**
Hard		25.3	**11.7**	37.9	**35.7**	35.9	28.8	33.8	39.5
Hard	✓	**31.1**	6.1	**44.0**	27.5	**58.8**	**32.7**	**43.1**	**47.2**

Semi+All strategy. Conversely, on EPIC-Kitchens-100 hard negatives are preferred, leading to 58.8% nDCG-SYN (+22.6%), 32.7% nDCG-BoW (+3.9%), 43.1% nDCG-PoS (+9.3%), and 47.2% mAP (+7.2%) by using HGR, and to 59.5% nDCG-SYN (+26.8%), 32.3% nDCG-BoW (+5.5%), 43.9% nDCG-PoS (+11.4%), and 45.1% mAP (+11.6%) by using Everything-at-once after a warm-up with semi-hard negatives. On MSVD the situation is skewed towards hard negatives in the case of HGR (achieving 44.0% nDCG compared to 37.9% of the baseline), and all negatives for Everything-at-once (41.2% nDCG compared to 35.1% of the baseline), although the difference with hard negative is small (40.4% nDCG). Such a different behavior may be a consequence of the distribution of the relevance values (Section 5.1), making harder negatives less informative in MSR-VTT. Furthermore, when utilizing the official MSR-VTT split, each video clip is paired to multiple captions, and the optimization process increases the similarity among all them. Thus, performing the optimization on all the negatives may pose an easier problem resulting in a model which generalizes better. Third, as mentioned before, the proposed strategy may lead on average to slightly lower recall rates (e.g., the average R@1 goes from 6.9% to 4.2% when using the *All* strategy in Everything-at-once): in fact, RANP assumes that multiple captions are relevant for the same video, and vice versa, thereby bringing them all closer in the embedding space (see Section 5.3 for further discussion and visualization). As a consequence, the recall rates may become lower because multiple elements which are unpaired to the query, yet semantically similar to it, may become closer to the anchor than the groundtruth.

5.2.3. Is a large scale instance-based pretraining helpful in SSVR?

Large scale pretraining is often performed to leverage transfer learning and, possibly, to ease the training process. Shvetsova et al. (2022) pretrained their proposed model on HowTo100M (Miech et al., 2019), a large scale dataset of tutorial clips. By leveraging the pretrained weights for Everything-at-once shared by the authors, we explore the effects of such a technique in the SSVR task on the official split of MSR-VTT.

According to the results shown in Table 2, starting the training process from the pretrained weights has a positive effect on both nDCG and recall rates in most of the cases. For instance, with the NCE loss originally used by the model, it leads to +2.2% nDCG-SYN (from 26.1% to 28.3%), +2.9% nDCG-BoW (from 37.1% to 40.0%), +2.9% nDCG-PoS (from 35.4% to 38.3%), and +1.2% R@1 (9.3% to 10.5%). Replacing the NCE with our version, i.e., NCE-RANP, we observe a further +2.8% nDCG-SYN (28.3% to 31.1%), +1.9% nDCG-BoW (from 40.0% to 41.9%), +2.5% nDCG-PoS (from 38.3% to 40.8%), but lower R@1 (from 10.5% to 7.9%). When applying the triplet loss to finetune the pretrained model, similar observations can be made, as consistent improvements are observed across the *All* and *Semi+All* strategies. Specifically, a state-of-the-art nDCG-SYN of 35.6% is obtained by using the *Semi+All* strategy and Triplet-RANP. Better nDCG-BoW (+1.6%) and nDCG-PoS (+1.5%) are obtained in a similar setup (*Semi+All* and Triplet-RANP) without the pretrain.

5.2.4. How does the training loss and the validation performance behave?

In Fig. 7, we analyze and compare the behavior of the standard Triplet loss function (green) to the proposed Triplet-RANP (red). We consider four cases: (above, left) HGR on EPIC-Kitchens-100; (above, right) EAO on EPIC-Kitchens-100; (below, left) HGR on MSR-VTT; (below, right) EAO on MSR-VTT.

Both the Triplet loss and Triplet-RANP follow a similar trend. Notably, the loss value for RANP is around twice the standard triplet because our loss consists of the sum of two terms (Eq. (9)).

The performance trend reported on the validation set shows that with Triplet-RANP, a considerable improvement is consistently observed in nDCG. In particular, for EPIC-Kitchens-100, the validation set consists of a small subset of the training set, as in previous works, e.g., Damen et al. (2021a). Therefore, the results in the first row of Fig. 7 confirm that with the proposed strategy it is possible to learn a function capable of ranking. Conversely, with the standard Triplet loss the quality of the ranked lists, measured with nDCG, is far lower. These results are confirmed in MSR-VTT (second row of Fig. 7), for which the validation set is separated from the training data, hence providing evidence of far better generalization.

5.2.5. Ablation study on the loss components

In Table 3, we present an ablation study on the loss components: we compare the proposed *Triplet-RANP* (i.e., Eq. (9)); *Triplet-RAN*, that is Eq. (9) without L_p, i.e., additional relevance-aware positives are not mined; *Triplet-RAP*, i.e., Eq. (9) only uses L_p; and the standard *Triplet* loss. The study is done on MSR-VTT using HGR (Table 3.a) and Everything-at-once (Table 3.b), serving respectively as a single-modal (appearance-only) and multi-modal (appearance and motion) model. For the former, we use the *Hard* strategy, whereas *All* was used for the latter. The results provide empirical evidence that improving the selection of the negatives by using RAN already leads to higher quality ranking lists, obtaining +3.4% nDCG-SYN (25.3% to 28.7%), +2.3% nDCG-BoW (36.1% to 38.4%), and +2.5% nDCG-PoS (34.0% to 36.5%) in HGR and +2.7% nDCG-SYN (24.8% to 27.5%), +2.3% nDCG-BoW (34.9% to 37.2%), and +2.3% nDCG-PoS (33.2% to

Table 2
Influence of HowTo100M-pretrain on Everything-at-once (Shvetsova et al., 2022) and subsequent finetuning with several strategies. Experiments performed on the official split of MSR-VTT.

PT	Opt	NCE-RANP	nDCG-SYN (%)	nDCG-BoW (%)	nDCG-PoS (%)	Mean R@1 (%)
✓	*Zero-shot*		21.5	31.9	30.6	9.2
	All		26.1	37.1	35.4	9.3
✓	All		28.3	40.0	38.3	**10.5**
	All	✓	28.7	39.9	38.7	4.4
✓	All	✓	31.1	41.9	40.8	7.9
PT	Opt	Triplet-RANP	nDCG-SYN (%)	nDCG-BoW (%)	nDCG-PoS (%)	Mean R@1 (%)
	All		24.8	34.9	33.2	6.9
✓	All		23.1	32.8	30.9	8.6
	All	✓	33.6	42.5	42.2	4.3
✓	All	✓	34.4	44.7	44.2	5.5
	Semi+All		21.3	30.7	29.1	8.2
✓	Semi+All		24.0	33.8	32.0	9.4
	Semi+All	✓	34.4	**45.3**	**44.8**	6.7
✓	Semi+All	✓	**35.6**	43.7	43.3	6.2

Table 3
Ablation study on MSR-VTT using **(a)** single-modal (appearance-only) and **(b)** multi-modal (appearance and motion) model. Discussion in Section 5.2.5.

(a) Appearance-only	nDCG-SYN			nDCG-BoW			nDCG-PoS			R@1		
Model	t2v	v2t	avg	t2v	v2t	avg	t2v	v2t	avg	t2v	v2t	avg
HGR	24.6	26.1	25.3	36.4	35.8	36.1	33.9	34.2	34.0	**9.4**	**14.1**	**11.7**
HGR+Triplet-RAN	27.4	30.1	28.7	38.9	37.9	38.4	36.4	36.6	36.5	8.9	13.0	10.9
HGR+Triplet-RAP	20.1	**34.1**	27.1	29.1	**48.5**	38.8	28.2	**46.8**	37.5	1.2	1.6	1.4
HGR+Triplet-RANP	**28.3**	33.8	**31.1**	**40.4**	45.6	**43.0**	**38.2**	44.6	**41.4**	5.5	6.6	6.1
(b) Multi-modal	nDCG-SYN			nDCG-BoW			nDCG-PoS			R@1		
EAO	23.9	25.6	24.8	35.0	34.7	34.9	33.0	33.3	33.2	6.4	**7.4**	6.9
EAO+Triplet-RAN	26.4	28.6	27.5	37.3	37.1	37.2	35.3	35.8	35.5	**7.0**	6.9	**7.0**
EAO+Triplet-RAP	28.8	33.5	31.1	40.7	**47.1**	**43.9**	39.3	**45.6**	**42.4**	1.4	0.7	1.0
EAO+Triplet-RANP	**31.5**	**35.7**	**33.6**	**41.3**	43.6	42.5	**40.4**	44.0	42.2	4.0	4.4	4.3

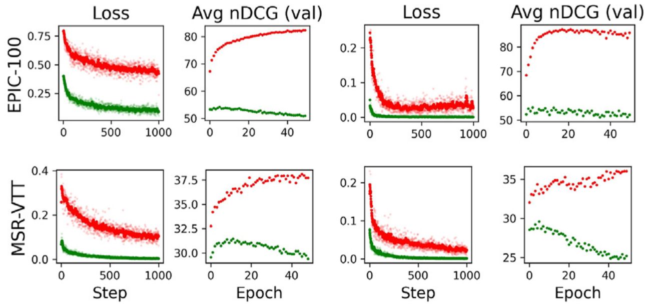


Fig. 7. Plot of the training loss (green: triplet loss, red: our Triplet-RANP), the average nDCG on the validation set, both for HGR (left) and EAO (right), on both datasets (above: EPIC-Kitchens-100, below: MSR-VTT). Note that Triplet-RANP is a sum of two losses, hence the absolute loss values are greater. Discussion in Section 5.2.4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

35.5%) in Everything-at-once. With Triplet-RAP, that is using only the relevance-aware positives and ignoring the groundtruth associations, an improvement of +1.8% (25.3% to 27.1%) and +6.3% nDCG-SYN (24.8% to 31.1%) are observed on the two networks. While similar improvements are seen in HGR (36.1% to 38.8% nDCG-BoW and 34.0% to 37.5% nDCG-PoS), in EAO using only the mined positives leads to better nDCG-BoW (34.9% to 43.9%) than using the full Triplet-RANP (obtaining 42.5% nDCG-BoW). Nonetheless, in the other metrics the proposed strategy obtains more favorable results. Notably, mining the additional positives has an important effect on the recall rates (e.g., going from 6.9% to 1.0% in Everything-at-once). This change is a direct consequence of bringing non-groundtruth relevant elements (i.e., the $q+$ selected by Eq. (6)) closer to the anchor at every iteration, which has a positive effect on SSVR performance, yet a negative one in IVR performance, further confirming the gap between the two tasks. Finally, by using the full Triplet-RANP, further improvements are achieved compared to both its components, leading to +5.8% nDCG-SYN, +7.9% nDCG-BoW, and +7.4% nDCG-PoS over HGR and +8.8% nDCG-SYN, +7.6% nDCG-BoW, and +9.0% nDCG-PoS over Everything-at-once. Note that similar results are obtained while using other strategies, e.g. +7.7% nDCG-SYN is achieved by HGR with RANP using the *All* strategy (see Table 1).

5.2.6. *Comparison with state-of-the-art*

In Tables 4, 5, and 6, we report the results obtained with our strategies and compare them to other methods, on EPIC-Kitchens-100, MSR-VTT, MSVD, and Charades. Given that SSVR was introduced recently, there are only a few works which perform the evaluation using nDCG and mAP on these datasets (Damen et al., 2021a,b, 2022; Falcon et al., 2022c; Wray et al., 2021). Therefore, we compare our results to them and report the results from other methods focusing on IVR performance.

EPIC-Kitchens-100. In Table 4, we compare to MME and JPoSE, proposed by Wray et al. (2019) and used in Damen et al. (2021a) as the baselines for the challenge. We include Hao et al. from the 2021 edition (Damen et al., 2021b), IIE-MRG (Damen et al., 2022), UniUD-UB-UniBZ (Falcon et al., 2022a), and Ego-VLP (Lin et al., 2022) from the 2022 edition of the EPIC-Kitchens-100 Challenge (Damen et al., 2022). Moreover, we also include the method proposed in Falcon et al. (2022c). It needs to be noted that the two methods which currently and jointly hold the state-of-the-art, i.e., UniUD-UB-UniBZ and Ego-VLP, are not fairly comparable to our method, since the former uses an ensemble of several methods, including HGR trained with RANP, and the latter performs an additional pretraining with more than 3 millions of egocentric clips. Therefore, we compare to the current single-model state-of-the-art, that is the method proposed in Falcon et al. (2022c). Both on HGR and Everything-at-once, the addition of RANP leads to considerable improvements: HGR+Triplet-RANP obtains around +2.6% (58.8% compared to 56.2%) nDCG and +1.4% mAP

Table 4
Comparison on state-of-the-art methods for EPIC-Kitchens-100. In-depth discussion at Section 5.2.6. Ego-VLP and UniUD-UB-UniBZ are not fairly comparable to our method, since the former uses a huge amount of additional egocentric data (3 millions annotated clips), and the latter uses an ensemble of models, including HGR+RANP.

EPIC-Kitchens-100	nDCG (%)			mAP (%)		
Model	t2v	v2t	avg	t2v	v2t	avg
HGR (Chen et al., 2020b)	37.9	41.2	39.5	35.7	36.1	35.9
EAO (Shvetsova et al., 2022)	35.2	37.3	36.2	33.9	40.8	37.4
MME (Wray et al., 2019)	46.9	50.0	48.5	34.0	43.0	38.5
JPoSE (Wray et al., 2019)	51.5	55.5	53.5	38.1	49.9	44.0
Damen et al. (2021b)	51.8	55.3	53.5	38.5	50.0	44.2
IIE-MRG (Damen et al., 2022)	54.1	56.6	55.3	38.1	47.5	42.8
Falcon et al. (2022c)	–	–	56.2	–	–	45.8
Ours						
EAO+NCE-RANP	55.7	58.5	57.1	38.7	44.2	41.4
EAO+Triplet-RANP	**57.5**	**61.6**	**59.5**	39.6	50.6	45.1
HGR+Triplet-RANP	56.5	61.2	58.8	**42.3**	**52.0**	**47.2**
Ego-VLP (Lin et al., 2022)	59.6	63.3	61.4	41.0	53.9	47.4
UniUD-UB-UniBZ (Falcon et al., 2022a)	58.9	63.2	61.0	44.4	55.2	49.8

Table 5
Comparison on several state-of-the-art methods for the official split of MSR-VTT. 'Num. mod.': number of modalities used for training. 'PT': pretrain on HowTo100M for EAO, or the OpenAI checkpoint for CLIP.

MSR-VTT			nDCG (%)			R@1 (%)	GMR (%)
Model	Num. mod.	PT	t2v	v2t	avg	avg	t2v
VSE (Mithun et al., 2018)	1		–	–	–	6.3	12.6
VSE++ (Mithun et al., 2018)	1		–	–	–	7.9	13.4
Multi Cues (Mithun et al., 2018)	3		–	–	–	9.7	16.3
MDMMT (Dzabraev et al., 2021)	3		–	–	–	–	41.4
CLIP-straight (Portillo-Quintero et al., 2021)	1	✓	–	–	–	30.8	35.4
Clip2Video (Fang et al., 2021)	1	✓	–	–	–	42.2	47.8
CLIP2TV (Gao et al., 2021)	1	✓	–	–	–	–	48.9
LAFF-ml (Hu et al., 2022)	4+4[a]		–	–	–	–	47.2
DVTR (Zhang et al., 2023a)	1	✓	–	–	–	43.5	50.7
RVTR (Zhang et al., 2023b)	1	✓	–	–	–	32.6	40.7
CE	1		28.9	30.0	29.4	5.9	12.7
MoEE	1		28.4	29.5	29.0	6.0	13.2
HGR	1		24.6	26.1	25.3	11.8	21.6
CE	7		32.2	32.9	32.6	13.1	22.9
MoEE	7		33.3	32.3	32.8	12.8	22.7
EAO	2		25.6	26.7	26.1	9.3	19.0
EAO	2	✓	27.8	28.8	28.3	10.5	21.3
CLIP4Clip	1	✓	29.2	30.6	29.9	42.5	49.8
ProST	1	✓	30.2	31.1	30.7	**43.3**	**51.8**
Ours							
HGR+Triplet-RANP	1		**33.0**	**37.8**	**35.4**	7.5	15.0
EAO+Triplet-RANP	2		32.5	36.3	34.4	6.6	13.6
EAO+Triplet-RANP	2	✓	**33.5**	**37.8**	**35.6**	6.2	13.7
CLIP4Clip+NCE-RANP	1	✓	**29.9**	**31.8**	**30.8**	38.2	43.4
ProST+NCE-RANP	1	✓	30.1	**32.0**	**31.1**	42.6	49.2

[a] LAFF-ml uses four visual experts for motion features and four textual experts for textual features.

(47.2% compared to 45.8%); on Everything-at-once the improvement measures up to +3.3% (59.5%) nDCG, yet the previous state-of-the-art maintains a small margin of +0.7% mAP. The comparison between our two RANP-trained methods shows that Everything-at-once leads to higher nDCG (59.5% vs 58.8%), whereas HGR achieves higher mAP (47.2% vs 45.1%), meaning that the latter allows to retrieve more highly relevant captions and videos to the top of the ranking list. This may be due to the hierarchical learning aspect of HGR, which can be quite important in EPIC-Kitchens-100 considering the structure of the available captions. Finally, in our experiments, CLIP4Clip did not converge properly on EPIC-Kitchens-100, hence we did not include it in the table.

MSR-VTT. For MSR-VTT, we compare to VSE, VSE++, and Multi Cues (Mithun et al., 2018), CLIP-straight (Portillo-Quintero et al., 2021), MDMMT (Dzabraev et al., 2021), Clip2Video (Fang et al., 2021), CLIP2TV (Gao et al., 2021), LAFF (Hu et al., 2022), RVTR (Zhang et al., 2023b), and DVTR (Zhang et al., 2023a) which report results on the official split of MSR-VTT. We also reproduce and report results with MoEE (Miech et al., 2018), CE (Liu et al., 2019), HGR, Everything-at-once, ProST, and CLIP4Clip. We report for each of these models the amount of modalities used, since each differs in this regards. We evaluated CE and MoEE both using only appearance features and using all the seven available modalities within the open source codebase of Liu et al. (2019). Both these models, even by using one modality, achieve higher scores (29.0% and 29.4%) than HGR and Everything-at-once (25.3% and 26.1%): considering that the latter two perform better in IVR, this further shows that SSVR requires different tools and strategies. This is further confirmed by the results obtained with CLIP4Clip, for which we used the open source codebase of Luo et al. (2022). In fact, the performance of this model in IVR is far better than the other models (49.8 GMR, compared to 22.9 obtained by CE with 7 modalities). However, when tested under the SSVR setting, it performs similarly to them (e.g., it obtains 29.9% average nDCG, whereas MoEE obtained 29.0%). The results change when adding RANP to the training. If HGR and Everything-at-once are trained with Triplet-RANP considerable improvements are observed, respectively achieving

Table 6
Comparison on several state-of-the art methods on MSVD. 'PT': pretrain on OpenAI checkpoint for CLIP.

MSVD			nDCG (%)			R@1 (%)	GMR (%)
Model	Num. mod.	PT	t2v	v2t	avg	avg	t2v
VSE (Mithun et al., 2018)	1		–	–	–	14.0	25.0
VSE++ (Mithun et al., 2018)	1		–	–	–	18.3	31.8
Multi Cues (Mithun et al., 2018)	2		–	–	–	25.9	39.0
CLIP-straight (Portillo-Quintero et al., 2021)	1	✓	–	–	–	48.4	55.9
Clip2Video (Fang et al., 2021)	1	✓	–	–	–	52.8	67.7
CLIP2TV (Gao et al., 2021)	1	✓	–	–	–	–	67.4
LAFF-ml (Hu et al., 2022)	4+4[a]		–	–	–	–	62.8
DVTR (Zhang et al., 2023a)	1	✓	–	–	–	60.5	71.1
HGR	1		37.7	38.2	37.9	35.7	54.0
EAO	2		35.4	34.8	35.1	31.7	53.9
CLIP4Clip	1	✓	39.4	35.5	37.4	**56.4**	**71.2**
ProST	1	✓	39.3	34.9	37.1	52.5	70.7
Ours							
HGR+Triplet-RANP	1		**41.4**	**46.7**	**44.0**	27.5	42.7
EAO+Triplet-RANP	2		41.0	41.4	41.2	25.6	47.0
CLIP4Clip+NCE-RANP	1	✓	38.3	37.8	38.0	54.2	66.5
ProST+NCE-RANP	1	✓	39.2	37.5	38.4	54.1	68.8

[a] LAFF-ml uses four visual experts for motion features and four textual experts for textual features.

35.4% (+10.1%) and 34.4% (+9.3%) nDCG, which are better than the nDCG obtained by CE and MoEE trained with seven modalities (32.6% and 32.8%). Although the pretrain helps Everything-at-once achieve 35.6% nDCG, it is not as fair to be compared with the other models which are not pretrained on large scale datasets. An improvement of 0.9% nDCG is also observed on CLIP4Clip, reaching 30.8% nDCG. In the case of ProST, the improvement in average nDCG measures +0.4%, with an improvement of +0.9% in terms of video-to-text nDCG. As mentioned in the previous sections, it can be again observed that using RANP, the SSVR performance generally increases at the cost of a lower IVR performance (e.g., in CLIP4Clip, the average R@1 goes from 42.5 to 38.2).

MSVD. As in the case of MSR-VTT, for MSVD we compare to VSE, VSE++, and Multi Cues (Mithun et al., 2018), CLIP-straight (Portillo-Quintero et al., 2021), Clip2Video (Fang et al., 2021), CLIP2TV (Gao et al., 2021), LAFF (Hu et al., 2022), and DVTR (Zhang et al., 2023a). We reproduce and report results with HGR, Everything-at-once, CLIP4Clip, and ProST. Table 6 shows that CLIP4Clip and ProST achieve very good results among the baselines, both in terms of nDCG (with CLIP4Clip achieving 37.4% average nDCG) and recall rates (with CLIP4Clip achieving 56.4% average R@1 and 71.2% GMR). Using the proposed training strategy, most of the methods considered in our analysis achieve considerable improvements, putting HGR as the top method in SSVR performance, obtaining 44.0% nDCG (+6.6% compared to CLIP4Clip). Notably, when combining NCE-RANP with ProST, an improvement is obtained both in IVR (+1.6% average R@1) and SSVR performance (+1.3% average nDCG). Nevertheless, as in previous cases, the other methods under analysis report lower IVR performance when combined with the RANP strategy. This is due to multiple samples becoming relevant and being pulled together in the embedding space, making it harder to discriminate the groundtruth.

Charades. Finally, we performed a comparison among the methods under analysis on Charades, used in this paper for action retrieval. We report results with HGR (*All*), Everything-at-once (*Semi+hard*), CLIP4Clip, and ProST, obtained with and without the proposed training strategy. Table 7 shows that ProST achieves the highest nDCG (89.2% on average) among the baselines, whereas CLIP4Clip and EAO achieve better recall rates (21.9% and 21.7% average R@1), with EAO achieving higher GMR (43.4%) due to higher R@5. It can be seen that in this dataset all the methods achieve fairly high nDCG, likely due to the distribution of relevance being skewed towards very small values (Fig. 5.d). This makes it harder to achieve significant improvements in the quality of the ranking lists. Nonetheless, with the proposed strategy, all the methods achieve moderate improvements in SSVR performance, ranging from +0.2% to +0.8% nDCG. As in previous experiments, using the proposed strategy leads to lower IVR performance.

Table 7
Comparison on Charades used for action retrieval. 'PT': pretrain on OpenAI checkpoint for CLIP.

Charades			nDCG (%)			R@1 (%)	GMR (%)
Model	Num. mod.	PT	t2v	v2t	avg	avg	t2v
HGR	1		88.4	89.4	88.9	16.6	28.7
EAO	2		88.5	89.3	88.9	21.7	**43.4**
CLIP4Clip	1	✓	88.3	89.5	88.9	**21.9**	34.5
ProST	1	✓	88.7	89.6	89.2	20.7	34.8
Ours							
HGR+Triplet-RANP	1		88.7	**90.7**	**89.7**	9.9	20.5
EAO+Triplet-RANP	2		88.2	89.6	88.9	15.1	30.1
CLIP4Clip+NCE-RANP	1	✓	88.3	89.8	89.1	16.6	27.3
ProST+NCE-RANP	1	✓	**89.0**	90.1	89.5	16.2	28.2

5.3. Qualitative analysis

In this section, several qualitative analyses are performed to provide insights into how the proposed strategy shapes the model towards better SSVR performance. Specifically, Sections 5.3.1 and 5.3.2 provide an overview on the interpretation of the performance metrics and their impact on the ranking list. Then, a few failure modes are highlighted in Section 5.3.3. Finally, an in-depth analysis of the learned joint embedding space provides further evidence and a clear explanation on the effectiveness of the proposed strategy for the SSVR task (Section 5.3.4).

5.3.1. Visualizing the full ranking lists

Fig. 8 shows the full ranking list of the 9668 clips produced by the four models (HGR and Everything-at-once, trained with or without Triplet-RANP) for three different queries on the test set of EPIC-Kitchens-100. By training with RANP, the ranking lists produced on the test set have most of the relevant videos at the top of the ranking list, e.g., it can be clearly seen in the first two queries, "wipe counter" and "put down bins". Conversely, the two models trained with the IVR methodology have all the relevant videos scattered across the whole ranking list, making their retrieval difficult. In the third query, "put tablecloth into cupboard", it can be observed that more relevant videos are retrieved among the top ranks in the models trained with Triplet-RANP than in the two models trained without it. Nonetheless, some highly relevant videos have high ranks, e.g., in HGR+RANP there is a highly relevant video near the middle of the list.

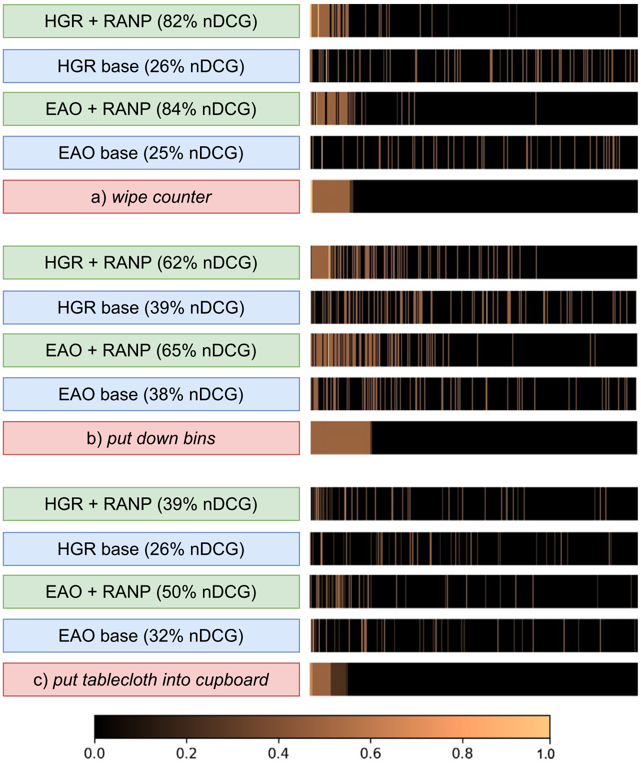


Fig. 8. Qualitative examples of text-to-video retrieval from the testing set. The full ranking lists (of length 9668) are shown and the color represents the relevance to the query (color scale shown below). The query is shown in red, along the optimal ranking lists. Both HGR and Everything-at-once display a similar behavior in bringing many more relevant videos to the top of the ranking list. More details in Section 5.3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.3.2. *Visualizing the top of the ranking lists*

Fig. 9 presents a more detailed visualization of the top 5 videos retrieved with three queries on the test set of EPIC-Kitchens-100. By looking at the query “continue wiping sink”, Fig. 9.a shows that HGR trained both with or without Triplet-RANP, and Everything-at-once trained with Triplet-RANP are able to retrieve highly relevant examples. In particular, the fifth video retrieved by “HGR base” (which is also the third retrieved by “EAO+RANP”) has similar appearance features but a relevance of 0.50 because it is captioned by “wipe off kitchen”, making its noun class unrelated to the more precise “sink”. Conversely, without RANP, Everything-at-once retrieves some videos which have similar motion but less precise appearance features (second, fourth, and fifth are not about “sinks”). In Figs. 9.b and 9.c the lists are more varied, although displaying some advantages in models trained with Triplet-RANP, such as in Fig. 9.b where both HGR and Everything-at-once are able to retrieve more highly relevant videos in the top five. It is interesting to observe how training with Triplet-RANP may lead to ranking lists in which the clips have a moderate relevance, despite not sharing the visual features: for instance, in Fig. 9.c “HGR+RANP” retrieves clips about “cutting” (same verb), whereas without RANP it looks for “bags” which are “opened”, though not “cut” through.

5.3.3. *Analysis of failure modes*

In Fig. 10, we added two examples of failure modes observed on both HGR and EAO. The first failure mode consists of partially relevant videos which are retrieved earlier than fully relevant ones, similar to what happens in Fig. 9.c. In fact, all the top 5 videos retrieved by “HGR+RANP” and by “EAO+RANP” in Fig. 10.a display actions comprising the correct verb (“open”; “cut”) but a wrong object (“fridge”, “cupboard”, etc instead of “container”; “cheese”, “vegetables”, etc instead of “beef”). The proposed strategy maximizes the similarity of all the samples whose relevance is higher than the threshold, but does not enforce any particular order among them: therefore, partially relevant elements may become more similar to the query than fully relevant ones. This is also confirmed by the embedding space analysis made later in this section. The second failure mode we observe is partially due to the complexity of the videos, which makes it difficult to annotate every relevant aspect of it. In fact, Fig. 10.b shows that, for the query “put lid”, “HGR+RANP” retrieves two videos whose caption contains additional information (“on pot” or “on pan”), therefore resulting in a relevance lower than 1. Similarly, for the query “hold pan”, the first two videos retrieved by “EAO+RANP” depict someone performing this action while another action takes place, resulting in partially irrelevant annotations (the first video is annotated by “shake saucepan”, whereas the second by “take pan”).

5.3.4. *Analysis of the joint embedding space*

Finally, we explore how the learned embedding space changes when using the proposed strategy. Specifically, Fig. 11 shows a subset of the embedding space (26000 random videos and captions from the training set) learnt on EPIC-Kitchens-100 by HGR trained with the standard triplet loss (Fig. 11.a) and with proposed strategy (Fig. 11.b). The same subset is used for both Figures. Each action is shown as a string containing its verb and noun class, e.g., “*verb_class-noun_class*”.

Using the standard methodology (Fig. 11.a), a cluster per action is obtained, with only small problems (e.g., an instance of “6-0” is far to the left, whereas all the other are clustered together on the right). However, the main shortcoming is that there are no semantic relations with neighboring clusters. For instance, close to the “4-12” cluster (bottom, left) there are no other clusters for verb 4 (“close”) or for noun 12 (“fridge”), since most of the neighbors deal with “putting {something} on” (verb 1), “cutting” (verb 7), “eggs” (noun 53), or “lighter” (noun 130). Similarly, cluster “8-0” has no neighboring clusters for “turn off” (verb 8) or “tap” (noun 0).

Conversely, with the proposed strategy, the embedding space is organized differently (Fig. 11.b). For instance, close to “turn off tap” (“8-0”) there is “turn on tap” (“6-0”), and close to that there is “turn on kettle” (“6-44”). Similarly, to the left there are several actions sharing the same verb 0 (“take”) and different nouns (2 - “plate”, 6 - “lid”, 7 - “bowl”, etc.). By reducing the distance between the clusters of semantically similar actions, the proposed strategy is able to build a joint embedding space in which semantic retrieval is possible and effective, as also confirmed by the quantitative results presented in the manuscript, therefore confirming the initial hypotheses.

5.4. *Limitations and future directions*

In this section, we highlight some research directions which would benefit solutions for the SSVR problem.

As described in Section 4, the proposed strategy requires computing relevance values at training time, and consequently part-of-speech tags and semantic classes. In our work, these steps are not learned during training and therefore are done in a preprocessing phase. To avoid it, the first research direction consists in understanding how to use the recent advancements in NLP to automatically measure the relevance value. We experiment with two possible solutions, and evaluate them qualitatively, yet no conclusive statement can be made. The first solution uses average-pooled GloVe embeddings for each sentence, and the cosine similarity as a relevance proxy. We observe that it might be possible to identify semantically similar sentences: for instance, “take pizza” has a similarity of 82.8% and 87.3% to “pick up pizza” and “take slice of pizza”, respectively. However, similar values are also obtained for opposite sentences, e.g., the similarity of “take pizza” and “put down pizza” is 82.3%, whereas “pour water into the kettle” and “remove the water from the kettle” have a similarity of 84.6%. The second technique we consider uses recent pretrained language models (Wang

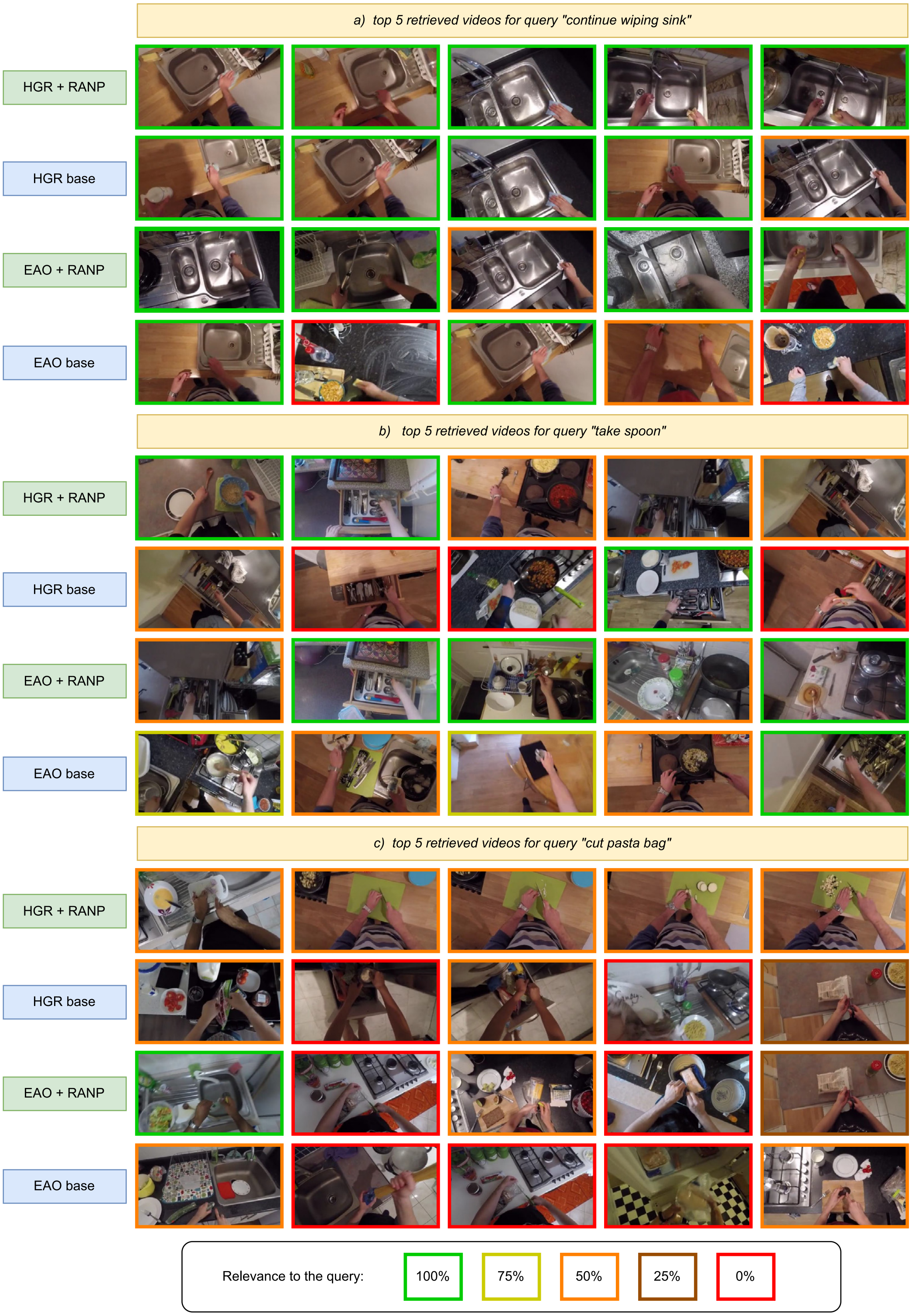

a) top 5 retrieved videos for query "continue wiping sink"
HGR + RANP
HGR base
EAO + RANP
EAO base
b) top 5 retrieved videos for query "take spoon"
HGR + RANP
HGR base
EAO + RANP
EAO base
c) top 5 retrieved videos for query "cut pasta bag"
HGR + RANP
HGR base
EAO + RANP
EAO base
Relevance to the query:
100%
75%
50%
25%
0%

Fig. 9. Qualitative examples of text-to-video retrieval from the testing set of EPIC-Kitchens-100. The border is colored green, yellow, orange, brown, or red based on the relevance to the query (respectively, 1.00, 0.75, 0.50, 0.25, 0.00). Discussion in Section 5.3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

et al., 2020). However, as in the previous case, this model is also unable to reliably distinguish semantically opposite sentences. For instance, given the sentence “take slice of pizza”, all the following variations (both similar and opposite) have very high similarity values: “pick up slice of pizza” (85.1%), “put slice of pizza somewhere” (86.0%), “put down slice of pizza” (81.9%), “throw slice of pizza” (80.4%), “drop slice of pizza” (80.1%). Therefore, differently from the technique used in our work, there is a need for further research to adapt these methods to reliably discriminate relevant from irrelevant sentences.

A second interesting direction emerges from the qualitative results discussed in Section 5.3. In fact, both the retrieval examples (Figs. 9 and 10) and the visualization of the embedding space (Fig. 11), clearly show that the proposed strategy rearranges the embedding space in a way which puts both fully and mildly relevant elements close. However, this means that for some queries partially relevant samples may be

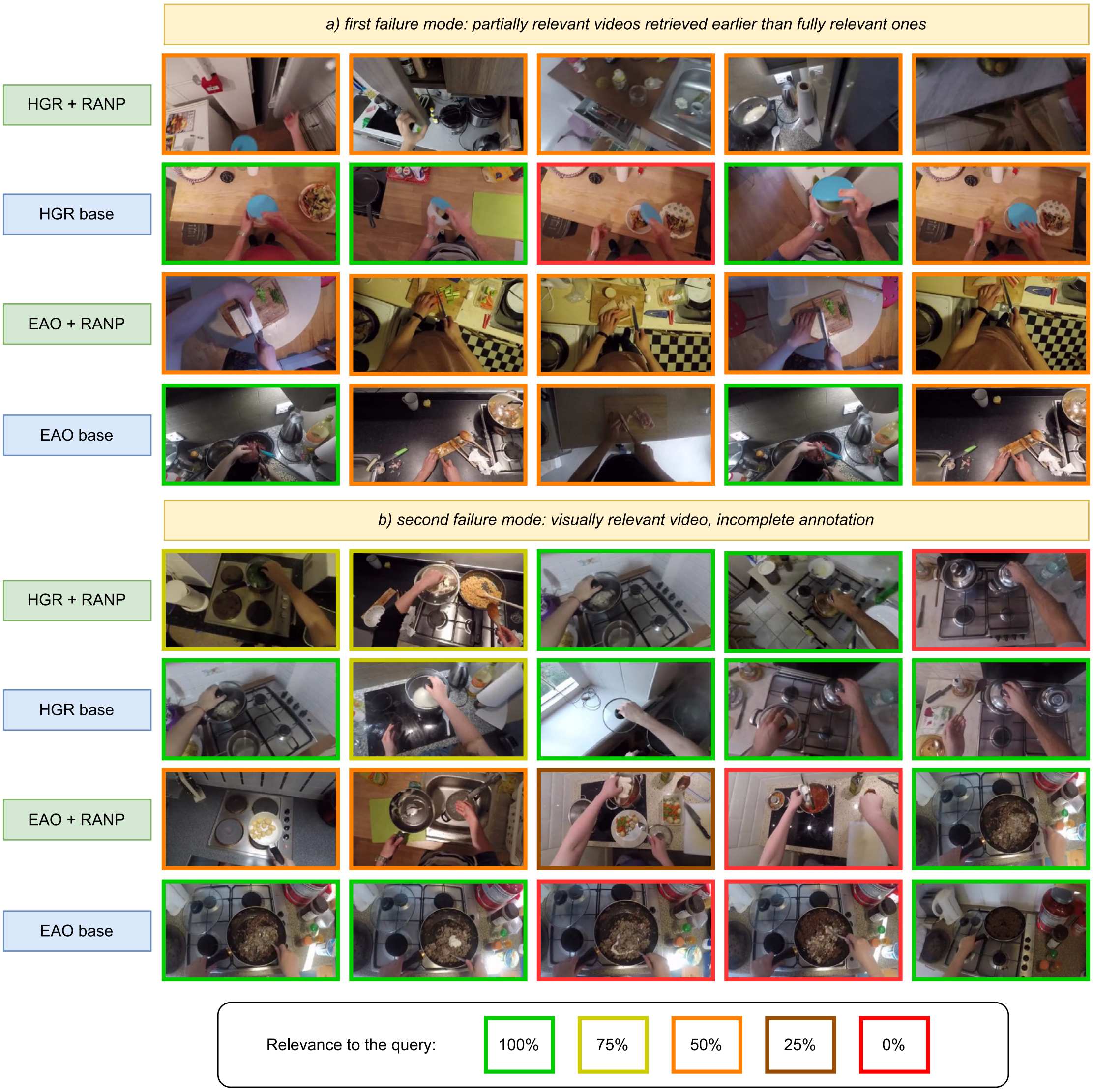

a) first failure mode: partially relevant videos retrieved earlier than fully relevant ones
HGR + RANP
HGR base
EAO + RANP
EAO base
b) second failure mode: visually relevant video, incomplete annotation
HGR + RANP
HGR base
EAO + RANP
EAO base
Relevance to the query:
100%
75%
50%
25%
0%

Fig. 10. Qualitative examples from the testing set of EPIC-Kitchens-100 of "failure modes" in models trained with the proposed strategy. The border is colored green, yellow, orange, brown, or red based on the relevance to the query (respectively, 1.00, 0.75, 0.50, 0.25, 0.00). Discussion in Section 5.3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

retrieved earlier than fully relevant samples. For instance, in the left area of Fig. 11.b, there is a "0-X" cluster where some "0-2" elements are closer to "0-6" than they are to the other "0-2" elements. This result is a consequence of the design of RANP, which considers all the videos and captions whose relevance is greater than τ as comparably relevant to the anchor element. Notably, it highly affects the IVR performance. Therefore, further research is required to design a methodology which simultaneously addresses both the IVR and the SSVR problem.

6. Conclusions

Recently, the community highlighted several limitations of Instance-based Retrieval (IVR), e.g., Chun et al. (2021), Wang et al. (2022) and Wray et al. (2021). In particular, a new problem was defined, namely the Semantic Similarity Video Retrieval problem (SSVR), which differently from IVR, it aims to retrieve *all* semantically equivalent videos for a given textual query, therefore leading to the need for highly different learning protocols and evaluation methodologies. So far, only few works attempted to tackle this problem, typically by using deep architectures originally designed for IVR and now customized for SSVR (Falcon et al., 2022c; Lin et al., 2022; Satar et al., 2022).

In this paper, we focused on customizing the behavior of IVR loss functions to better address SSVR. To do so, we started from state-of-the-art IVR methods, typically trained with a contrastive loss, e.g., the Triplet loss (Schroff et al., 2015) or the NCE loss (Miech et al., 2020). With these techniques, a neural network learns to output similar descriptors for each paired video and caption. Yet, they assume that all the other samples are completely irrelevant. We showed this assumption hardly holds in practice, and that it leads to suboptimal results for SSVR, due to the selection of negatives which share similar semantics as the query. Moreover, because only the video and caption pairs in the dataset are considered valid, there are many captions which could be used as positives for a video, but are not.

To address these two shortcomings, we proposed a novel strategy, using the overlap of semantic concepts between captions to improve the selection of the negative examples, while also discovering for each query new positives not originally paired to it in the dataset. We reformulated two popular loss functions based on our strategy, and tested them on four heterogeneous state-of-the-art IVR methods, both graph-based (Chen et al., 2020b) and Transformer-based (Luo et al., 2022; Shvetsova et al., 2022; Li et al., 2023). We validated our strategy on four datasets, EPIC-Kitchens-100, MSVD, Charades,

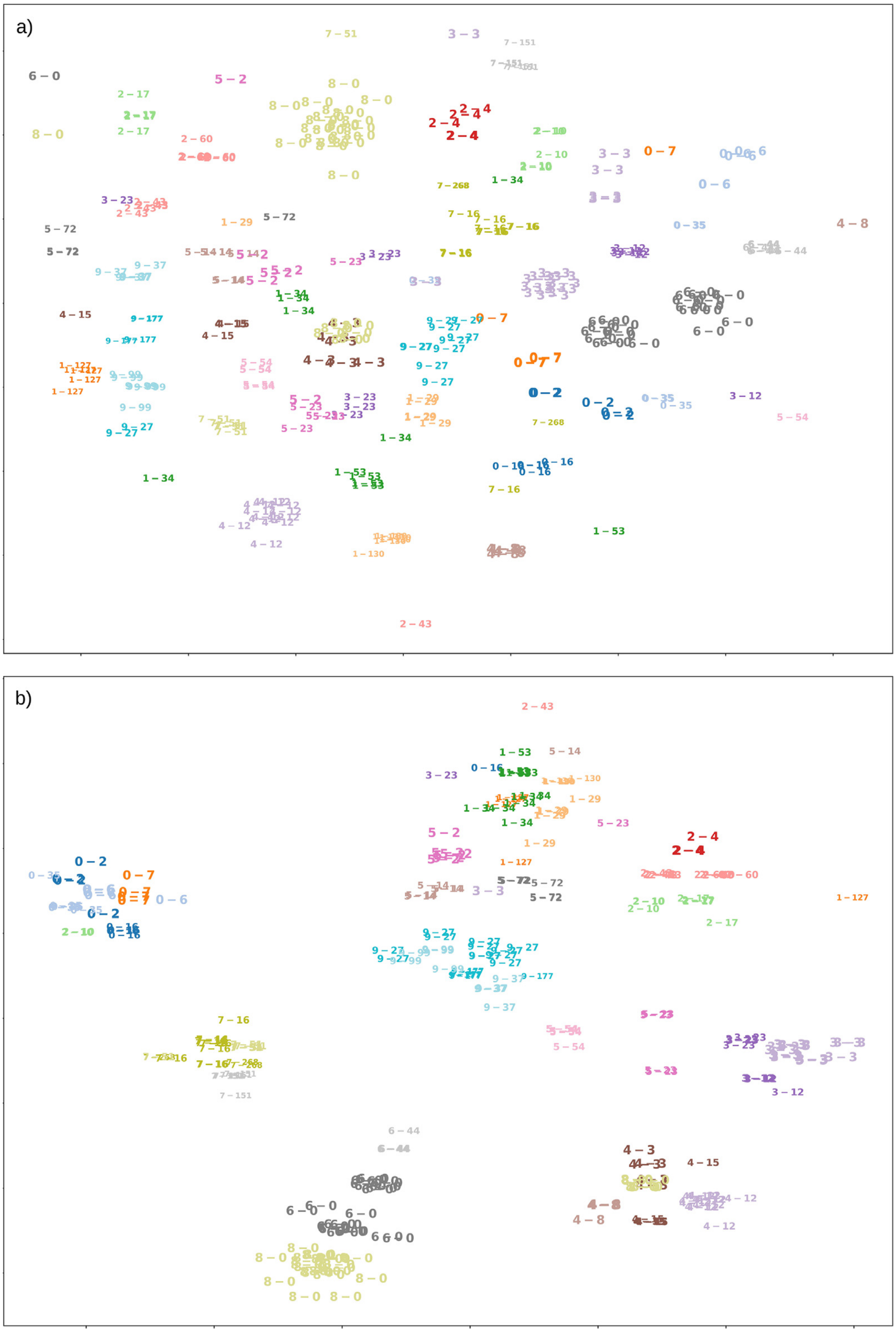


Fig. 11. t-SNE representation of the embedding spaces learned by HGR on EPIC-Kitchens-100. (**a**) HGR is trained with the proposed Triplet-RANP strategy. (**b**) HGR is trained with the standard triplet loss. Discussion in Section 5.3.

and MSR-VTT, and conducted an extensive quantitative analysis, comprising multiple experiments to provide evidence on the effectiveness of the proposed strategy to overcome the aforementioned limitations. Moreover, we obtained considerable improvements on previous state-of-the-art, e.g., +3.3% nDCG and +1.4% mAP on EPIC-Kitchens-100. The in-depth qualitative analyses further analyzed the impact of the proposed strategy on the quality of the ranked lists, and explained its effectiveness by reasoning on the structure of the learned embedding space. Finally, we highlighted the limitations of our approach and possible future directions.

CRediT authorship contribution statement

Alex Falcon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Giuseppe Serra:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition. **Oswald Lanz:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The video and text data which support the evaluation of this research study are available in the EPIC-Kitchens-100 (Damen et al., 2021a), MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), and Charades (Sigurdsson et al., 2016) datasets. For the first two datasets, the pre-extracted features can be found at https://github.com/epic-kitchens/C5-Multi-Instance-Retrieval and https://github.com/ninatu/everything_at_once respectively. In our repo, located at https://github.com/aranciokov/ranp/, we both provide the code and the pretrained models, as well as the data already packaged in the desired format.

Acknowledgments

This work was supported by MUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2022 (project code 2022YTE579), by the Department Strategic Plan (PSD) of the University of Udine – Interdepartmental Project on Artificial Intelligence (2020–25), and by Startup Grant IN2814 (2021–24) of University of Bolzano. 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. Open access funding provided by University of Udine within the CRUI-CARE Agreement.

References

Albanie, S., Liu, Y., Nagrani, A., Miech, A., Coto, E., Laptev, I., Sukthankar, R., Ghanem, B., Zisserman, A., Gabeur, V., et al., 2020. The end-of-end-to-end: A video understanding pentathlon challenge (2020). arXiv preprint arXiv:2008.00744.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proc. of IEEE CVPR, pp. 5297–5307.

Ashutosh, K., Girdhar, R., Torresani, L., Grauman, K., 2023. Hiervl: Learning hierarchical video-language embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23066–23078.

Baeza-Yates, R., Ribeiro-Neto, B., et al., 1999. Modern information retrieval, vol. 463, ACM press New York.

Chen, W., Chen, X., Zhang, J., Huang, K., 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proc. of IEEE CVPR. pp. 403–412.

Chen, D., Dolan, W.B., 2011. Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 190–200.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. In: Proc. of ICML, PMLR, pp. 1597–1607.

Chen, S., Zhao, Y., Jin, Q., Wu, Q., 2020b. Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proc. of IEEE/CVF CVPR. pp. 10638–10647.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: Proc. of IEEE CVPR. pp. 539–546.

Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D., 2021. Probabilistic embeddings for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8415–8424.

Croitoru, I., Bogolin, S.-V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y., 2021. Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proc. of IEEE/CVF ICCV. pp. 11583–11593.

Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al., 2021a. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. IJCV 130, 33–55.

Damen, D., Fragomeni, A., Munro, J., Perrett, T., Whettam, D., Wray, M., Furnari, A., Farinella, G.M., Moltisanti, D., 2021b. EPIC-KITCHENS-100- 2021 Challenges Report. Technical Report, University of Bristol.

Damen, D., Fragomeni, A., Perrett, T., Whettam, D., Wray, M., Zhu, B., Furnari, A., Farinella, G.M., Moltisanti, D., 2022. EPIC-KITCHENS-100- 2022 Challenges Report. Technical Report, University of Bristol.

Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M., 2021. Dual encoding for video retrieval by text. IEEE TPAMI 44 (8), 4065–4080.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A., 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proc. of IEEE/CVF CVPR, pp. 3354–3363.

Falcon, A., Serra, G., Escalera, S., Lanz, O., 2022a. Uniud-FBK-UB-UniBZ submission to the EPIC-kitchens-100 multi-instance retrieval challenge 2022. arXiv preprint arXiv:2206.10903.

Falcon, A., Serra, G., Lanz, O., 2022b. Learning video retrieval models with relevance-aware online mining. In: Proc. of ICIAP. pp. 182–194.

Falcon, A., Sudhakaran, S., Serra, G., Escalera, S., Lanz, O., 2022c. Relevance-based margin for contrastively-trained video retrieval models. In: Proc. of ICMR.

Fang, H., Xiong, P., Xu, L., Chen, Y., 2021. Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097.

Gabeur, V., Sun, C., Alahari, K., Schmid, C., 2020. Multi-modal transformer for video retrieval. In: Proc. of IEEE ECCV. Springer, pp. 20020–20029.

Gao, Z., Liu, J., Sun, W., Chen, S., Chang, D., Zhao, L., 2021. Clip2tv: Align, match and distill for video-text retrieval. arXiv preprint arXiv:2111.05610.

Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., Luo, P., 2022. Bridging video-text retrieval with multiple choice questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16167–16176.

Gutmann, M., Hyvärinen, A., 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proc. of Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 297–304.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: Proc. of IEEE CVPR. pp. 1735–1742.

Harwood, B., Kumar B.G., V., Carneiro, G., Reid, I., Drummond, T., 2017. Smart mining for deep metric learning. In: Proc. of IEEE ICCV. pp. 2821–2829.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proc. of IEEE/CVF CVPR. pp. 9729–9738.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737.

Hu, F., Chen, A., Wang, Z., Zhou, F., Dong, J., Li, X., 2022. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In: European Conference on Computer Vision. Springer, pp. 444–461.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. ACM TOIS 20 (4), 422–446.

Jiang, Q.Y., He, Y., Li, G., Lin, J., Li, L., Li, W.J., 2019. SVD: A large-scale short video dataset for near-duplicate video retrieval. In: Proc. of IEEE/CVF ICCV, pp. 5281–5289.

Lesk, M., 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proc. of SIGDOC. pp. 24–26.

Li, D., Li, J., Li, H., Niebles, J.C., Hoi, S.C., 2022. Align and prompt: Video-and-language pre-training with entity prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4953–4963.

Li, P., Xie, C.-W., Zhao, L., Xie, H., Ge, J., Zheng, Y., Zhao, D., Zhang, Y., 2023. Progressive spatio-temporal prototype matching for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4100–4110.

Lin, K.Q., Wang, A.J., Yan, R., Xu, E.Z., Tu, R., Zhu, Y., Zhao, W., Kong, W., Cai, C., Wang, H., et al., 2022. Egocentric video-language pretraining@ EPIC-KITCHENS-100 multi-instance retrieval challenge 2022. arXiv preprint arXiv:2207.01334.

Liu, Y., Albanie, S., Nagrani, A., Zisserman, A., 2019. Use what you have: Video retrieval using representations from collaborative experts. In: Proc. of BMVC.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3202–3211.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T., 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing 508, 293–304.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A., 2020. End-to-end learning of visual representations from uncurated instructional videos. In: Proc. of IEEE/CVF CVPR. pp. 9879–9889.

Miech, A., Laptev, I., Sivic, J., 2018. Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., Sivic, J., 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proc. of IEEE/CVF ICCV. pp. 2630–2640.

Miller, G.A., 1995. WordNet: a lexical database for english. Commun. ACM 38 (11), 39–41.

Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K., 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proc. of ACM ICMR. pp. 19–27.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W., 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proc. of IEEE/CVF CVPR. pp. 11205–11214.

Portillo-Quintero, J.A., Ortiz-Bayliss, J.C., Terashima-Marín, H., 2021. A straightforward framework for video retrieval using clip. In: Mexican Conference on Pattern Recognition. Springer, pp. 3–12.

Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., Cui, Y., 2021. Spatiotemporal contrastive video representation learning. In: Proc. of IEEE/CVF CVPR. pp. 6964–6974.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Satar, B., Zhu, H., Zhang, H., Lim, J.H., 2022. Exploiting semantic role contextualized video features for multi-instance text-video retrieval EPIC-KITCHENS-100 multi-instance retrieval challenge 2022. arXiv preprint arXiv:2206.14381.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: Proc. of IEEE CVPR. pp. 815–823.

Shvetsova, N., Chen, B., Rouditchenko, A., Thomas, S., Kingsbury, B., Feris, R.S., Harwath, D., Glass, J., Kuehne, H., 2022. Everything at once-multi-modal fusion transformer for video retrieval. In: Proc. of IEEE/CVF CVPR. pp. 20020–20029.

Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A., 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 510–526.

Sohn, K., 2016. Improved deep metric learning with multi-class n-pair loss objective. In: Proc. of NeurIPS. pp. 1857–1865.

Soviany, P., Ionescu, R.T., Rota, P., Sebe, N., 2022. Curriculum learning: A survey. Int. J. Comput. Vis. 130 (6), 1526–1565.

Suh, Y., Han, B., Kim, W., Lee, K.M., 2019. Stochastic class-based hard example mining for deep metric learning. In: Proc. of IEEE/CVF CVPR. pp. 7251–7259.

Wang, Z., Gao, Z., Xu, X., Luo, Y., Yang, Y., Shen, H.T., 2022. Point to rectangle matching for image text retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4977–4986.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M., 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv. Neural Inf. Process. Syst. 33, 5776–5788.

Wang, X., Zhu, L., Yang, Y., 2021. T2vlad: global-local sequence alignment for text-video retrieval. In: Proc. of IEEE/CVF CVPR. pp. 5079–5088.

Wray, M., Doughty, H., Damen, D., 2021. On semantic similarity in video retrieval. In: IEEE (Ed.), Proc. of IEEE/CVF CVPR. pp. 3650–3660.

Wray, M., Larlus, D., Csurka, G., Damen, D., 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In: Proc. of IEEE ICCV. pp. 450–459.

Wu, D., Li, H., Tang, Y., Guo, L., Liu, H., 2022. Global-guided asymmetric attention network for image-text matching. Neurocomputing 481, 77–90.

Xu, J., Mei, T., Yao, T., Rui, Y., 2016. Msr-vtt: A large video description dataset for bridging video and language. In: Proc. of IEEE CVPR. pp. 5288–5296.

Xuan, H., Stylianou, A., Liu, X., Pless, R., 2020a. Hard negative examples are hard, but useful. In: Proc. of IEEE ECCV. pp. 126–142.

Xuan, H., Stylianou, A., Pless, R., 2020b. Improved embeddings with easy positive triplet mining. In: Proc. of IEEE/CVF WACV. pp. 2474–2482.

Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., Chua, T.-S., 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1339–1348.

Yang, X., Wang, S., Dong, J., Dong, J., Wang, M., Chua, T.-S., 2022. Video moment retrieval with cross-modal neural architecture search. IEEE Trans. Image Process. 31, 1204–1216.

Zhang, H., Yang, Y., Qi, F., Qian, S., Xu, C., 2023a. Debiased video-text retrieval via soft positive sample calibration. IEEE Trans. Circuits Syst. Video Technol..

Zhang, H., Yang, Y., Qi, F., Qian, S., Xu, C., 2023b. Robust video-text retrieval via noisy pair calibration. IEEE Trans. Multimed..

Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R., 2023. Learning video representations from large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6586–6597.

Zhao, Y., Zhao, L., Zhou, X., Wu, J., Chu, C.-T., Miao, H., Schroff, F., Adam, H., Liu, T., Gong, B., et al., 2024. Distilling vision-language models on millions of videos. arXiv preprint arXiv:2401.06129.