Leveraging Generative Language Models for Weakly Supervised Sentence Component Analysis in Video-Language Joint Learning

Zaber Ibn Abdul Hakim [1,*] Najibul Haque Sarker [1,*] Rahul Pratap Singh [2]
Bishmoy Paul [1] Ali Dabouei [3,†] Min Xu [3,†]

[1]Bangladesh University of Engineering and Technology
[2]Netaji Subhas University of Technology [3]Carnegie Mellon University

zaberhakim666@gmail.com nhsarker.bd@gmail.com rahulprsingh1@gmail.com
paul.bish98@gmail.com ad0046@mix.wvu.edu mxu1@cs.cmu.edu

Abstract

A thorough comprehension of textual data is a fundamental element in multi-modal video analysis tasks. However, recent works have shown that the current models do not achieve a comprehensive understanding of the textual data during the training for the target downstream tasks. Orthogonal to the previous approaches to this limitation, we postulate that understanding the significance of the sentence components according to the target task can potentially enhance the performance of the models. Hence, we utilize the knowledge of a pre-trained large language model (LLM) to generate text samples from the original ones, targeting specific sentence components. We propose a weakly supervised importance estimation module to compute the relative importance of the components and utilize them to improve different video-language tasks. Through rigorous quantitative analysis, our proposed method exhibits significant improvement across several video-language tasks. In particular, our approach notably enhances video-text retrieval by a relative improvement of 8.3% in video-to-text and 1.4% in text-to-video retrieval over the baselines, in terms of R@1. Additionally, in video moment retrieval, average mAP shows a relative improvement ranging from 2.0% to 13.7 % across different baselines.

1. Introduction

With the rise of social media, streaming platforms, surveillance systems, and the entertainment industry, video has been solidified as one of the prime sources of information and recreation. Natural language is one of the most important modalities that accompanies video data due to its human-friendly and descriptive nature. Inclusion of both modalities in a unified task requires the interaction between videos and user input texts to interact in a multi-modal domain, giving rise to video-language joint learning tasks that include Video Retrieval [8, 15, 32], Video Captioning [17, 55], Action Segmentation [4, 12], Video Moment Retrieval [23, 28, 34], Video Summarization [13], etc. The

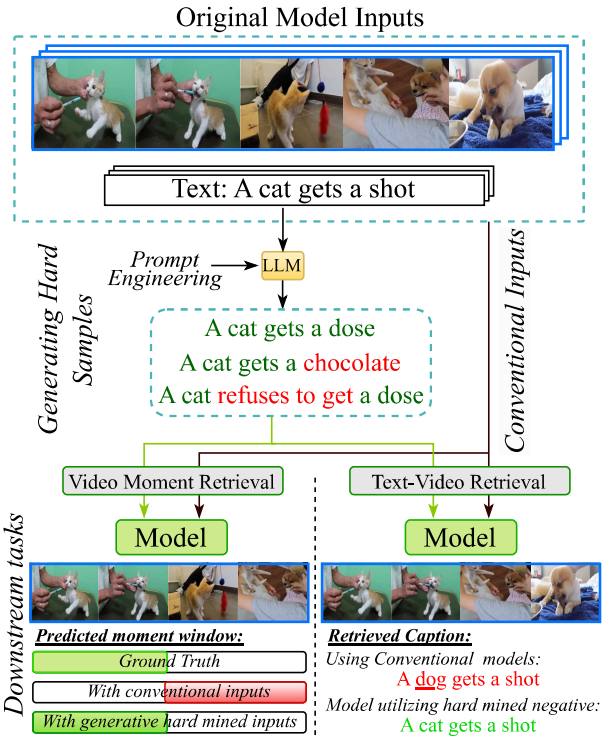

Figure 1. Performance comparison between existing approaches and our proposed method in video-language joint tasks. Here, conventional models fail to properly attend all components of the sentence, *e.g.* ***shot***. This is alleviated by the incorporation of targeted hard samples using our proposed mechanism.

*Equal Contribution
†Corresponding Author

Anchor Text (A)	type	Generated Samples (M)	Similarity between A & M
Young tourist couple sharing some videos of their tour.	Negative	Young tourist couple sharing some videos of their wedding.	0.906
	Positive	Some videos of their tour are being shared by the young tourist couple.	0.881

Table 1. Example of negative and positive samples generated using LLMs. By utilizing the capabilities of the LLMs we can generate very hard samples. It is supported by the generated negative samples being much closer to anchor text in embedding space compared to the corresponding positive text.

performance of models in these tasks depends on their capability to extract video features and align them with corresponding text features. Attending properly to each component of a sentence and assigning them the appropriate level of significance with regard to the video modality is an important requirement of these models. Our exploratory analysis suggests that the current models fall short in attending to all components appropriately which results in suboptimal outputs. One such instance is shown in Figure 1.

Here in moment-retrieval, the baseline model shows bias towards the subject *'cat'* and fails to connect it with the object *'shot'*. A parallel situation occurs in video-to-text retrieval, where the model struggles to distinguish that the subject *'dog'* in the retrieved text is unrelated to the object *'shot'*. Consequently, although both the models' predictions include the subject, the object is absent. This problem stems from the models' difficulty in representing all sentence components while aligning with video features.

The limited perception of textual features in video-language joint learning tasks can be rooted back to the presence of noisy text labels in web-crawled image-caption pairs [11, 22] used in the pre-training stage of encoders such as CLIP [36]. Correcting the huge amount of text labels in the pre-training stage is a significant challenge, motivating numerous efforts [10, 16, 25, 49, 52] to improve the textual representation in many downstream tasks with limited dataset. A common strategy is to introduce extra weak labels, aiding models in distinguishing sentence differences and ultimately enhancing feature representation. In earlier works [16, 52, 58], such additional negative and positive samples have been mined from the original dataset based on the similarity of the text representations. While the results are promising, the procedure for generating additional samples presented in those works is not controllable and there is no way to emphasize specific sentence parts.

A popular solution among researchers to generate additional labels in a more controllable manner involves leveraging the vast knowledge of a pre-trained large language model (LLM). In the majority of the cases [11, 22, 25, 49], LLM has been used to generate descriptive and refined information from the videos or the text labels. Apart from this, in some concurrent works [7, 33], LLM has also been used to introduce completely new positive or negative samples to be used in a contrastive learning setting. Although [33] emphasized the significance of verbs in video-language joint learning, a comprehensive investigation into various sentence components, such as objects, subjects, adjectives, etc., and their relative importance in understanding video-text correlation remains unexplored. In this work, we leverage LLM to generate hard negative samples from the original (anchor) samples that emphasize different sentence parts. This involves using precise prompt engineering to modify specific parts of the sentence while keeping the rest unchanged. Additionally, by completely restructuring the sentence, we create positive samples that lie relatively far from the anchor in the embedding space, compared to their negative counterparts, as shown in Table 1. We incorporate these generated samples using a modified contrastive loss function that considers the relative importance of different sentence parts. A visual comparison between an existing approach and our approach has been illustrated in Figure 1.

To sum up, the main contributions of our work are:

- We devise a mechanism for generating hard negative and positive samples for video-text joint learning tasks that emphasize different sentence components.
- We propose a pipeline that utilizes the generated samples to evaluate the importance of different sentence components for the computation of adaptive contrastive loss.
- Through extensive quantitative evaluations on two major video-text joint learning tasks, we demonstrate consistent performance improvement from the baseline in both scenarios. Furthermore, we conduct qualitative investigations to provide insights into the models' decision-making process after the integration of our approach.

2. Related Works

Multi-modal video-language joint learning tasks such as moment retrieval and video retrieval aim to establish meaningful alignment between the embedding spaces of textual and visual modalities [43, 45]. Recent advancements in the field have leveraged pre-trained encoders, *e.g.*, CLIP [36], as the feature extractor. They also use contrastive loss to align global or fine-grained representations of the text and video embeddings. Models utilizing global representations [15, 32] typically focus on `[CLS]` token for text and joint representation of frame-level embeddings. In contrast, models employing fine-grained representations [21, 24, 34] delve into the alignment of word-level representation of text with frame-level representations of video.

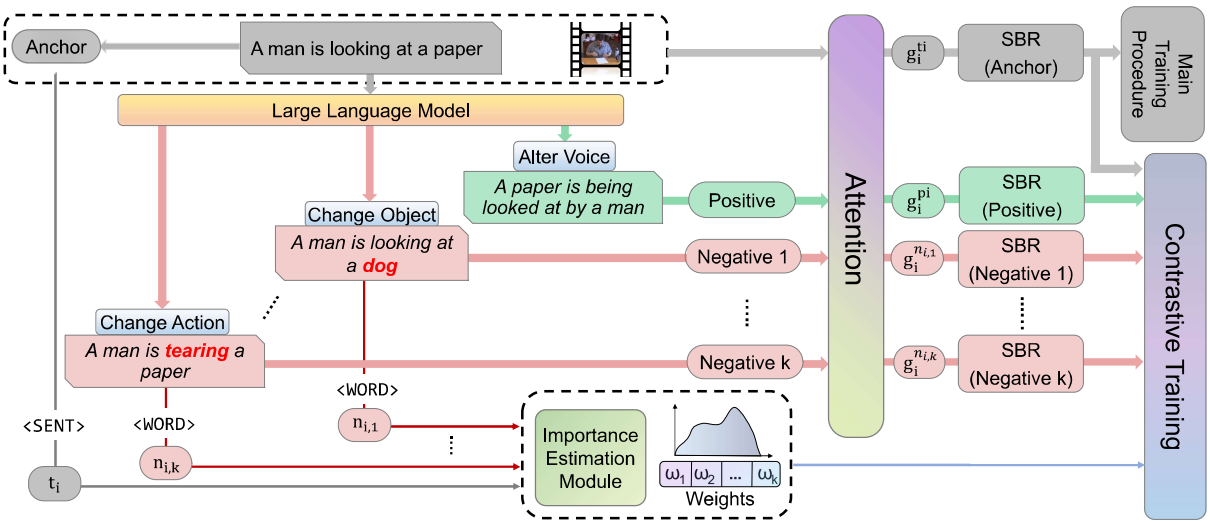

Figure 2. An overview of our proposed method. The videos, original or anchor texts, generated positive texts, and negative texts are passed through the Attention mechanism to generate Similarity-Based Representations, denoted as SBR in the figure (The exact method of generation varies with different tasks). The Contrastive Training procedure utilizes the generated texts to improve video-language context understanding. The Importance Estimation Module uses both the sentence and word representations (denoted as <SENT> and <WORD> respectively) to assign a weight to the losses associated with each of the negative samples.

2.1. Contrastive Learning

Contrastive learning is based on minimizing the distance between similar semantic features (positive pairs) while maximizing the distance between dissimilar ones (negative pairs) [18]. In moment retrieval, [31] utilizes contrastive loss where video-text pairs are positive samples if they belong in the training data. To align video and textual representations, [35] uses target moment clips as positive examples and other clips in that video as negative examples in a dual contrastive setting. While [46] mines hard positive video moments using query similarity across different videos, [56] generates easy and hard negative moments from the same video using a learnable Gaussian mask.

2.2. Additional Sample Generation

The usage of additional text samples, specifically hard-negative samples, has been explored in various literature [2, 6, 9, 39, 44]. Existing methods of hard-negative sampling are based on extracted features or using metadata and supervision labels [19, 20, 38, 41]. In the context of multi-modal tasks, [52] uses a cascading sampling method to mine video-text pairs as negative samples in a batch, in contrast to the random sampling done by [29, 57]. [16] utilizes the adaptive margin mechanism [40] for video retrieval as a better alternative to random sampling. [10] introduces Relevance-Aware Negatives and Positives mining (RANP) which uses semantics to mine samples. On the other hand, [58] excludes strongly connected negative samples by using embeddings to select better negative samples. [51] generates hard negatives through nearest neighbor retrieval. Recently the potential of Large Language Models (LLMs) has been explored in additional sample mining. Few works [25, 49] explored LLMs by generating additional comprehensive details from the available modalities, such as image captions or videos. [7] generates positive samples via LLM prompt engineering and hard negative samples by first masking sentence parts and then unmasking using LLM.

3. Methodology

To address the existing models' limitations in correlating sentence parts with suitable video representations, we present a method for generating challenging negative and positive samples targeting specific sentence parts. These samples improve the perception of specific sentence components, eventually increasing the understanding of video-language correlation. We use the generated samples as auxiliary samples alongside the original training samples by employing a novel adaptive contrastive loss. The proposed approach, summarized in Figure 2, is application-agnostic and can be adopted successfully in any video-language task.

This section is organized as follows: We outline our sample generation procedure in Subsection 3.1. Followed by this, Subsection 3.2 describes the procedure of applying our proposed contrastive loss by incorporating generated samples. In Subsection 3.3, we introduce our importance estimation module that adaptively weighs different sentence components based on their saliency. Finally, Subsection 3.4 provides the details of different video-language tasks on which we evaluate our method.

3.1. Sample Generation

Let $t_1, \ldots, t_m$ be the text samples available in a video-language joint learning task, where m is the total number of text samples in the training set. This textual information is the input to the conventional models. As previously discussed, a major shortcoming of these models is that they do not attend to all information in the sentence well enough. To force the models to have a better understanding of the cor-

relation of different sentence parts with videos, we generate hard negative and positive samples focusing on these components which are later employed in the training. We use a pre-trained LLM to generate such samples by utilizing their huge prior knowledge of the language.

Given the anchor text t_i, we generate k hard negative texts, $n_{i,1}, n_{i,2}, \ldots, n_{i,k}$, and a positive text, p_i, by leveraging the linguistic capability of a pre-trained LLM. For the generation of negatives, we instruct the LLM to specifically change a sentence part, *i.e.* verb, object, subject, etc. For the majority of the cases, the LLM only modifies the targeted part of the anchor text, keeping the rest of the sentence the same. Conversely, when generating the positive sample, we instruct the LLM to generate a sample that has a completely different sentence structure from the anchor while maintaining its semantics. We formalize the process of generating a negative and positive sample as:

$$\begin{aligned} n_{i,j} &\leftarrow \texttt{LLM}(t_i, \text{“Change} < part >_j \text{ of the sentence”}), \\ p_i &\leftarrow \texttt{LLM}(t_i, \text{“Alter voice of the sentence”}), \end{aligned} \tag{1}$$

where $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, k\}$, and the variable “$part$” represents any of the k sentence parts which are modified to generate the negative sample. Finally, we use CLIP's [37] text encoder to generate text embeddings from different types of texts. In our formulation, we have used similar notations ($\mathbf{t}, \mathbf{n}, \mathbf{p}$) to denote both texts and text embeddings interchangeably.

3.2. Incorporating the Generated Samples

The generated hard negative and positive samples force the existing models to discern the distinction among different words for a specific sentence part and their association with the video. This improves the overall perception of the video-language correlation. To incorporate these additional samples, along with the model's original samples, we compute a contrastive loss from the new samples and use it in association with the model's original loss. We use the embeddings of the three types (anchors, negatives, and positives) of texts mentioned in the preceding subsection, or their composite embedding with the videos to compute the contrastive loss. This approach facilitates the effective utilization of the generated additional auxiliary samples.

Let $\mathbf{g}_i$ be a general embedding for i^{th} sample, which can either be text embeddings or video-text composite embeddings. Given three types of text embeddings ($\mathbf{t}_i, \mathbf{n}_{i,j}, \mathbf{p}_i$), the general embedding for anchor texts, negative texts, and positive texts are denoted with $\mathbf{g}_i^{t_i}, \mathbf{g}_i^{n_{i,j}}$, and $\mathbf{g}_i^{p_i}$ correspondingly. Then the contrastive loss for i^{th} sample is formulated as following:

$$\mathcal{L}_i = -\log \frac{e^{\text{sim}(\mathbf{g}_i^{t_i}, \mathbf{g}_i^{p_i})/\tau}}{e^{\text{sim}(\mathbf{g}_i^{t_i}, \mathbf{g}_i^{p_i})/\tau} + \sum_{\mathbf{n} \in \mathbf{S}_i} e^{\text{sim}(\mathbf{g}_i^{t_i}, \mathbf{g}_i^{n})/\tau}}, \tag{2}$$

where τ is the temperature coefficient, $\mathbf{S}_i$ is the set of embeddings of all negative texts for i^{th} sample, and $\text{sim}(\cdot, \cdot)$ represents the function to compute the similarity between the two embeddings.

3.3. Weakly Supervised Importance Estimation of Sentence Components

Although we generate multiple types of negative texts by modifying specific components, *e.g.* noun, verb, or object, different sentence components don't exert similar importance in context understanding. For example, objects can play a more important role in some sentences whereas in a different example, it might be irrelevant. The aforementioned concern makes it important to evaluate the relative importance of each type of component adaptively for each sample. In the contrastive loss, defined in Equation 2, all the negatives are treated similarly. Considering this, we adaptively choose the most discernable sentence component per instance instead of using all of them together. The positive effect of this strategy highlights the importance of finding an optimal combination of all sentence components. To address this, we introduce a weakly-supervised sentence component analysis module. This module adaptively predicts the saliency of sentence components for each anchor text without any direct supervision. These dynamically computed scores denote which sentence components are crucial in context understanding, as depicted in Figure 4.

Using the Most Discernable Component. As we are using completely unsupervised sample generation, some generated samples might be completely unrelated to the anchor sentence. For instance, when a sentence doesn't have any adjective, using an adjective-changed negative won't make much sense. To address this issue, we focus only on the most discernible component.

To achieve this, we decompose the general contrastive loss of the i^{th} sample, $\mathcal{L}_i$, into k contrastive losses ($\mathcal{L}_{i,1}, \ldots, \mathcal{L}_{i,k}$), each corresponding to a specific negative. We compute them with a slight modification to the formulation presented in Equation 2. Instead of using a single set of negative text embeddings ($\mathbf{S}_i = \{\mathbf{n}_{i,1}, \ldots \mathbf{n}_{i,k}\}$), we compute the losses using k different sets of negatives, each containing one type of negative sample. The minimum of these decomposed losses is related to the component that the model can identify most confidently. So, we compute the loss for i^{th} sample as below:

$$\mathcal{L}_i = \min(\mathcal{L}_{i,1}, \mathcal{L}_{i,2}, \ldots, \mathcal{L}_{i,k}) \tag{3}$$

Computation and utilization of Importance Weight Estimation. This module utilizes a cross-attention mechanism to attend each type of negative text embedding with the anchor text embedding. We aim to consider the association between the sentence representation of anchor text embedding, $\mathbf{st}_i$, with all of its corresponding word-level representations of negative texts, $\mathbf{n}_{(i,:)}$, to generate a weight for each of the negatives. After applying the attention, we use

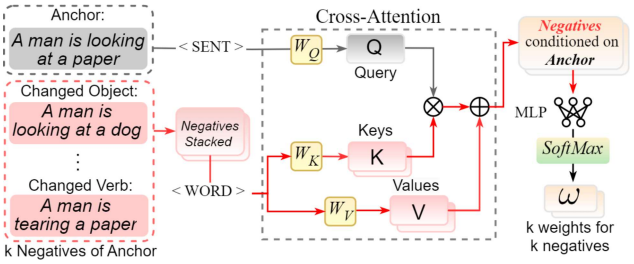

Figure 3. Importance Estimation Module. The sentence and word representations (denoted as <SENT> and <WORD> respectively) from Anchor and Negative text are passed through the cross-attention module to generate weights. The $\bigotimes$ operation denotes the dot products of the keys with the query followed by a softmax operation, and the $\bigoplus$ operation is the weighted averaging of all word-level tokens of values.

a linear transformation followed by a softmax activation to generate the weights.

Let dim be the vector dimension of each token in the text embeddings. $W_Q, W_K, W_V \in \mathbb{R}^{dim \times h}$ denote the learnable weight vectors for the linear transformations to transform the text representations into "query", "key", and "value" correspondingly, where h is the dimension of the hidden state. In addition, $W_\omega \in \mathbb{R}^h$ is the weight vector for the final linear transformation layer that is applied on the weighted "value" vector. The operations are formulated as:

$$\begin{gathered} Q_i = W_Q * \mathbf{st}_i, K_{(i,:)} = W_K * \mathbf{n}_{(i,:)}, V_{(i,:)} = W_V * \mathbf{n}_{(i,:)}, \\ V'_{(i,:)} = \texttt{Cross\_Attention}\left(Q_i, K_{(i,:)}, V_{(i,:)}\right), \\ \mathbf{m}_{(i,:)} = W_\omega * V'_{(i,:)}, \\ \boldsymbol{\omega}_{(i,:)} = \frac{e^{\mathbf{m}_{(i,:)}}}{\sum_j e^{\mathbf{m}_{(i,j)}}}, \end{gathered} \tag{4}$$

where $\boldsymbol{\omega}_{(i,:)} \in \mathbb{R}^k$ the estimated importance scores.

Finally, rather than using the simple contrastive loss, illustrated in Equation 2, we combine the component-wise contrastive losses, $\mathcal{L}_{i,1}, \ldots, \mathcal{L}_{i,k}$, as below:

$$\mathcal{L}_i = \sum_{j=1}^{k} \boldsymbol{\omega}_{i,j} * \mathcal{L}_{i,j}. \tag{5}$$

3.4. Attention Between Video and Text

For generating the multi-modal feature in video-language tasks, different baseline works adopted various approaches which is also reflected in the general definition of similarity computation in Equation 2. In a subset of works [5, 14, 32], simple cosine similarity between video and text embeddings is employed for such purpose. On the other hand, the majority of the works tend to use different variations of attention mechanisms [23, 26, 34]. Given this diversity, there is no fixed approach for all video-language joint learning tasks to generate this multi-modal feature. In this subsection, we discuss such differences and highlight how we adopt different baseline works into our approach.

3.4.1 Video Moment Retrieval

Generally, video moment retrieval works with video embeddings that have been attended by text embeddings. To achieve this, a multi-head self-attention layer [42, 48] or a cross-modal attention layer [1, 47, 54] has been used in recent works [23, 27, 34]. The self-attention approach concatenates the video and text embeddings on sequence dimension and then passes it to the conventional self-attention layer [23]. Conversely, in the case of cross-modal attention, videos are used as queries and texts are treated as keys and values for the attention mechanism[27, 34].

We formalize a generalizable formulation for the aforementioned attention mechanism that generates the joint embedding, $\mathbf{g}_i$, used in Equation 2 as below:

$$\mathbf{g}_i^{a_i} = \texttt{Attention}_{v-a}(\mathbf{v}_i, \mathbf{a}_i) \tag{6}$$

where $\mathbf{v}_i$ is the video embedding, $\mathbf{a}_i$ can be the text embedding of either anchor, negative, or positive texts of i^{th} sample. Then the $\texttt{sim}(\cdot,\cdot)$ function in Equation 2 is represented as a simple cosine similarity of the aforementioned joint embeddings ($\mathbf{g}_i^{a_i}$).

3.4.2 Video-Text Retrieval

Video-text retrieval tasks generally work by aligning the videos and their corresponding texts in a shared embedding space. While, initial studies [15, 30] typically focused on the global representations of texts and videos for computing similarity score, recent studies [21, 32, 53] have shown significant improvement by employing word-level and frame-level embeddings in addition to global representations. This enables better alignment of video frames with text words.

We formulate the procedure that generates the similarity score between any text pair as follows:

$$S_{t-a} = \texttt{Attention}_{t-a}(\mathbf{t}_i, \mathbf{a}_i), \tag{7}$$

where $\mathbf{t}_i$ is the word-level embedding of any anchor text, and $\mathbf{a}_i$ represents the same for any positive or negative texts of i^{th} sample. The details relating to the implementation of Function 7 are elaborated in Section 7 of the supplementary material. This function takes the place of the $\texttt{sim}(\cdot,\cdot)$ function in contrastive loss Equation 2.

4. Exeriments

Datasets. We use the QVHighlights [23] dataset for performance evaluation on moment retrieval task. It contains over 10,000 videos and each video includes relevant clips associated with human-written text queries. Additionally, we use the MSVD [3] dataset to evaluate performance on video and text retrieval tasks. It contains 1,970 videos and

	Moment-DETR [24]					QD-DETR [34]				
	R1		mAP			R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg	@0.5	@0.7	@0.5	@0.75	Avg
Baseline	52.89	33.02	54.82	29.40	30.73	62.40	44.98	62.52	39.88	39.86
Simple Contrastive Loss	55.32	36.45	56.88	32.88	33.75	61.80	45.01	62.23	40.39	40.58
Most Discernable	56.61	36.71	57.64	33.83	34.26	62.19	44.94	**62.62**	40.49	40.34
Adaptive	**56.81**	**37.42**	**58.30**	**33.95**	**34.94**	**63.04**	**45.85**	62.59	**41.29**	**40.66**

Table 2. Performance comparison of different settings on QVHighlights test set for moment retrieval. In the experiments other than baseline, all the negative samples have been used together.

80K captions, averaging ∼40 captions per video. We use the split proposed by Xu et al. [50].

Implementation Details. We experiment using 2 NVIDIA RTX 3090 24 GB GPUs using the Pytorch library. We use the default implementation of Moment-DETR [24] and QD-DETR [34] for moment retrieval and X-CLIP for video-text retrieval tasks. We train X-CLIP on a batch size of 30 with $6.1e^{-5}$ as the learning rate. Further details of our setup are provided in the Supplementary Material Section 8.

4.1. Performance Comparison

4.1.1 Moment Retrieval

Moment-DETR [24]. The results on the test set of QVHighlights for the Moment-DETR model are presented in Table 2. Here, we consider the non-pretrained version of the Moment-DETR model as the baseline. As conventional models are deficient in attending to different parts of the sentences relevant to the video, we utilize prompt engineering to create hard negative samples by targeting individual components of the sentence: verb, adjective/adverb, subject, and object. Additionally, we also convert the texts to negated passive, which generates samples with accentuated objects. Initially, we experiment with a simple contrastive loss configuration on positive and hard negative samples generated by LLM. This increases the model's performance from the baseline's average mAP score of 30.73 to 33.75, which is a relative improvement of 9.8%, suggesting that by using targeted hard negatives, the model learns to attend to all parts of the sentence. Subsequently, we devise a contrastive loss setting where instead of using all the generated negatives for a sample, we choose the one that is most discernable for that specific sample. The objective is to observe how the performance is affected when negatives are selected based on the input sample. This further outperforms the baseline average mAP by 11.48% and validates our assumption that for each sample text, there is an optimal combination of negatives which enables the model to put proper emphasis on the relevant part of the sentence. Consequently, after introducing our proposed methodology of weakly supervised adaptive importance estimation of sentence components, the score further improves to 34.94 for average mAP outperforming the baseline Moment-DETR model by 13.7%.

	T2V			V2T		
	R@1↑	R@5↑	MnR↓	R@1↑	R@5↑	MnR↓
XCLIP [32]	50.0	80.0	8.8	64.8	91.1	3.0
Simple CL	50.2	80.1	8.7	68.0	91.0	2.8
Discernable	50.0	80.0	8.7	67.9	92.2	3.0
Adaptive	**50.7**	**80.3**	**8.4**	**70.2**	**94.0**	**2.3**

Table 3. Comparison of Video-Text Retrieval performance on MSVD [3] with batch size 30.

QD-DETR [34]. The findings for QD-DETR on the QVHighlights test set are outlined in Table 2 and it corroborates the observations made with Moment-DETR. Here, we consider the non-pretrained version of QD-DETR model as the baseline, which works with only video and text data. We use the same negatives used for the Moment-DETR model, initially incorporating them in the simple contrastive loss setting. This results in an average mAP score of 40.58 which is a 1.8% improvement from the baseline. For the most discernable setting, even though there is a minute decrease in average mAP, other metrics show improvement. This again supports the idea of introducing a mechanism for adaptive importance estimation of different parts of the sentence. Consequently, our attention-based adaptive importance mechanism provides an average mAP score of 40.66, which is a 2.0% improvement over the baseline.

4.1.2 Video-Text Retrieval

X-CLIP [32]. Table 3 outlines the result of X-CLIP model on the test set of MSVD [3] dataset. We generate hard negatives by changing the verb and the subject for this dataset. It can be observed that these hard negatives generated using LLM provide a significant improvement of 3.2% in R@1 score of the text retrieval task and a marginal improvement of 0.2% R@1 score of the video retrieval task just from simple contrastive loss. Furthermore, using weakly supervised importance estimation of sentence components, the model yields a significant improvement of 5.4%, 2.9%, and 0.7 in R@1, R@5, and Mean Rank respectively for Video-to-Text (V2T) retrieval task and 0.7%, 0.3% and 0.4 improvement of R@1, R@5 and Mean Rank for the Text-to-Video (T2V) retrieval task, over the baseline. The huge improvement in the V2T task can be attributed to the model's high sensitivity to the differences between sentences and its hard negative, since this property is especially pertinent in the

	Moment-DETR [24]					QD-DETR [34]				
	R1		mAP			R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg	@0.5	@0.7	@0.5	@0.75	Avg
Baseline	53.87	34.00	55.42	28.96	30.90	61.61	45.61	62.07	41.16	40.92
Negated Verb Query	**57.35**	39.55	**57.43**	33.69	34.45	**63.35**	**47.87**	**62.94**	**42.76**	**41.83**
Negated Adjective Query	55.94	38.13	57.12	34.04	35.10	61.74	46.45	62.09	40.92	41.10
Object Changed Query	55.94	40.39	57.39	35.51	**35.76**	61.94	46.26	62.42	41.45	41.24
Negated Passive Query	56.26	**41.29**	57.21	**35.75**	35.63	62.32	47.03	61.91	42.59	41.62
Changed Subject Query	56.84	38.06	57.37	33.98	34.28	63.10	46.39	62.53	41.14	40.88

Table 4. Comparison of performance with baseline while using individual hard negatives on moment retrieval models: Moment-DETR and QD-DETR. The reported scores are on the validation set of the QVHighlights dataset.

	T2V			V2T		
	R@1↑	R@5↑	MnR↓	R@1↑	R@5↑	MnR↓
XCLIP [32]	50.0	80.0	8.8	64.8	91.1	3.0
Verb	49.9	**80.1**	8.7	**66.8**	**93.2**	**2.9**
Subject	**50.2**	80.0	**8.6**	**66.8**	92.5	**2.9**

Table 5. Contribution of different hard negatives with X-CLIP baseline on MSVD [3] with batch size 30. 'Verb': changing verb, 'Subject': changing subject i.e. they represent the part of the caption that is modified to form a new caption.

V2T task, which evaluates all the texts in the dataset for their relevance to a single video. This further supports the hypothesis that introducing the weakly supervised adaptive importance estimation mechanism with generated hard negatives contributes to improved performance.

4.2. Ablation Study and Discussion

4.2.1 Impact of sample generation criteria

We perform ablation studies to present how each of the negative samples generated by Equation 1 affects the overall performance by using a default contrastive learning setting with a single type of negative and a single positive at a time.

Moment Retrieval. Table 4 illustrates the impact of individual hard negative on the outcomes of moment retrieval models on the QVHighlights validation set. For Moment-DETR [24], every type of negative improves the performance compared to the baseline. This suggests that generating negative samples using any criterion has notable potential for improving the performance of video-text tasks. We also observe that the object-changed negative queries result in the best overall average mAP. This suggests that the baseline model provides comparatively low attention to objects present in queries. The table also presents the results for the QD-DETR [34] model where improvement in performance for each of the different negatives can be observed as well. Here we observe that the most improvement comes by utilizing negated verbs. This highlights that the baseline model has a limited capability in correlating query verbs with the videos in moment retrieval task.

Video Retrieval. The results on the test set of the MSVD dataset, using individual negatives with X-CLIP [32] as the baseline, are presented in Table 5. Compared to baseline scores, notable improvement can be observed in the V2T metrics, with marginal improvements in T2V metrics as well. The results suggest that the baseline model in video-text retrieval has a limited capability in correlating both, query verbs and subjects with videos.

The increase in performance observed through these experiments substantiates our objective of generating automated hard negative samples utilizing LLMs by targeting specific parts of sentences in the text modality.

4.2.2 Effect of Contrastive training

The main objective of using contrastive loss to incorporate the generated positives and negatives is to force the model to perceive the distinction among various words for a specific part of the sentence. We can infer from Figure 5, that the base model puts the anchor text and negative texts very close in the embedding space shown by their higher similarity of sentence representation. The proximity in embedding space often makes the model misjudge between opposite samples. Whereas, after the inclusion of the generated samples using contrastive loss, the model learns to push the negative samples further away from the anchor in embedding space. This is indicated by the very low similarity score between these opposite types of texts after inclusion.

4.2.3 Effect of Importance Estimation mechanism

Our importance estimation module is based on the observation that the significance of a specific part of a sentence varies for different samples. A text has specific parts that are more relevant to the sample where the baseline model might not provide enough attention. To present that our proposed methodology is alleviating these deficiencies, we provide some samples comparing the outputs of the baseline models with proposed models in Figure 4. In the first sample, the baseline model appears to be ignoring the object *orange lei* in the query: "*Man in striped shirt is wearing a* orange lei". Our mechanism assigns the most importance to the respective object-changed hard negative sample, thus making the model attend to the object part of the query. The resultant

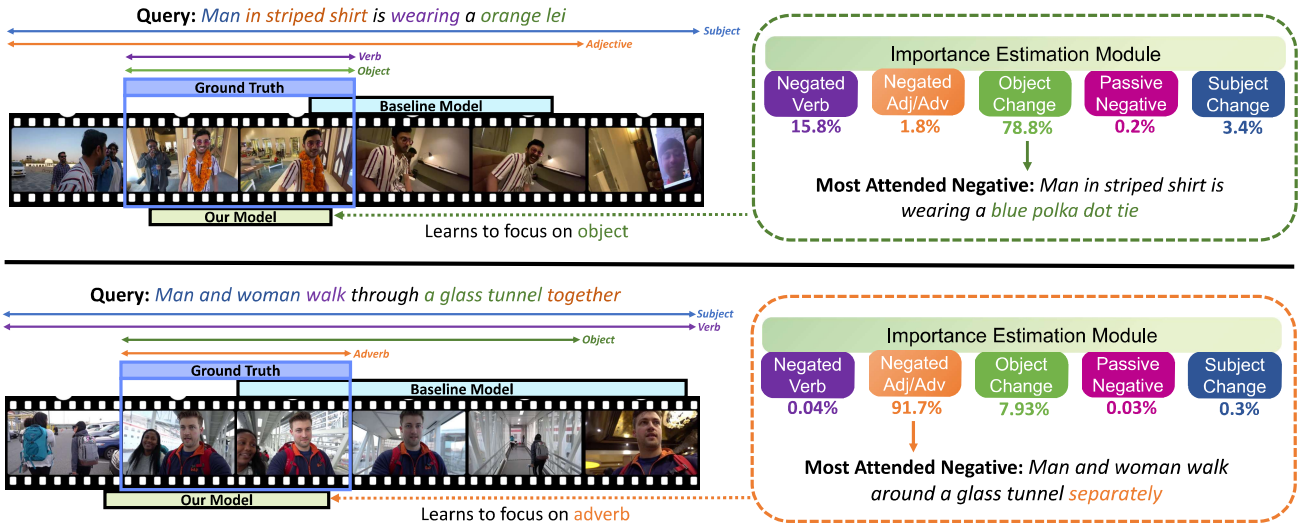

Figure 4. Examples of moment retrieval from baseline and our proposed version of Moment-DETR. The first example portrays a scenario where the original model struggled with the object *orange lei* in the query which is remedied by our importance estimation module giving more weight to object changed negative which an LLM generated by changing the object to *blue polka dot tie*. The second example provides a similar scenario where the model struggled with attending the adverb *together* but is then remedied by our mechanism providing more weight to the adverb changed negative which an LLM generated by changing the adverb to its antonym *separately*. In both cases, our mechanism learned to prioritize negative queries that address deficiencies in parts of the sentence where the baseline model struggled.

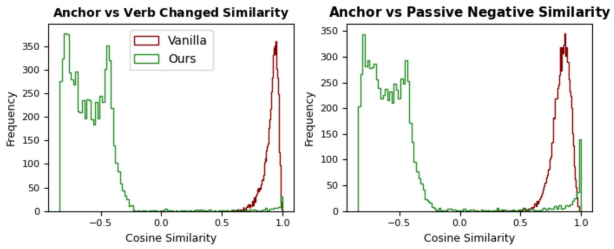

Figure 5. Distribution of similarity of sentence representation between "Anchor and Verb-changed Negatives", and "Anchor and Passive Negatives". Before applying our method, there were difficulties in discerning different types of texts, indicated by the high similarity of opposite texts. With our method, models effectively push negative text embeddings further from anchors, demonstrated by the low similarity of opposing texts.

model can successfully provide attention to the object resulting in better predictions. It is also noted that the second highest weighted negative is the negated verb query, which is relevant to the video and is also insignificantly attended by the baseline model.

While the baseline model provides less attention to the object in the first example, the second example illustrates an instance where the baseline model cannot discern the adverb *together* in the query: "*Man and woman walk through a glass tunnel* together". This further demonstrates the need for weakly supervised importance estimation for each video-text pair. We observe the proposed model assigning the highest weight to the respective negated adverb sample. Furthermore, in the video timeline, we observe that the baseline model also insufficiently recognizes the object *glass tunnel* in the video and the corresponding object-changed negative is the second highest weighted negative sample. Consequently, our proposed model can provide accurate predictions where all of the relevant parts of the query are represented.

5. Conclusion

We present a novel framework for weakly supervised adaptive contrastive learning for multi-modal video-language tasks. Specifically, we mitigate the issue of models' deficiency in the perception of different sentence components by utilizing LLM-generated component-targeted negative samples. We additionally integrate our proposed adaptive importance estimation module to accommodate sample-wise variations in the significance of different types of sentence components. We evaluate our method across three different baselines and two different video-language joint learning tasks. In each task, our method outperforms the baseline considerably, validating the effectiveness of our proposed method.

6. Acknowledgement

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629, DBI-2238093, IIS-2007595, IIS-2211597, and MCB-2205148. This work was supported in part by Oracle Cloud credits and related resources provided by Oracle for Research, and the computational resources support from AMD HPC Fund.

References

[1] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8127–8137, 2021. 5

[2] Rui Cao, Yihao Wang, Yuxin Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding. *arXiv preprint arXiv:2202.13093*, 2022. 3

[3] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 5, 6, 7

[4] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 1

[5] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 5

[6] Esra Dönmez, Pascal Tilli, Hsiu-Yu Yang, Ngoc Thang Vu, and Carina Silberer. Hnc: Leveraging hard negative captions towards models with fine-grained visual-linguistic comprehension capabilities. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 364–388, 2023. 3

[7] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 2, 3

[8] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 1

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 3

[10] Alex Falcon, Giuseppe Serra, and Oswald Lanz. Learning video retrieval models with relevance-aware online mining. In *International Conference on Image Analysis and Processing*, pages 182–194. Springer, 2022. 2, 3

[11] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023. 2

[12] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 1

[13] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1

[14] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 5

[15] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022. 1, 2, 5

[16] Feng He, Qi Wang, Zhifan Feng, Wenbin Jiang, Yajuan Lü, Yong Zhu, and Xiao Tan. Improving video retrieval by adaptive margin. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1359–1368, 2021. 2, 3

[17] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 1

[18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2, 2020. 3

[19] Qing-Yuan Jiang, Yi He, Gen Li, Jian Lin, Lei Li, and Wu-Jun Li. Svd: A large-scale short video dataset for near-duplicate video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5281–5289, 2019. 3

[20] Ruijie Jiang, Thuan Nguyen, Prakash Ishwar, and Shuchin Aeron. Supervised contrastive learning with hard negative samples. *arXiv preprint arXiv:2209.00078*, 2022. 3

[21] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 39–48. Association for Computing Machinery, 2020. 2, 5

[22] Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*, 2023. 2

[23] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 1, 5

[24] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2, 6, 7

[25] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. *arXiv preprint arXiv:2307.03135*, 2023. 2, 3

[26] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925, 2021. 5

[27] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 5

[28] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 1

[29] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3

[30] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022. 5

[31] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 156–171. Springer, 2020. 3

[32] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 1, 2, 5, 6, 7

[33] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 2

[34] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. *arXiv preprint arXiv:2303.13874*, 2023. 1, 2, 5, 6, 7

[35] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[38] Jinfeng Rao, Hua He, and Jimmy Lin. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1913–1916, 2016. 3

[39] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 3

[40] David Semedo and João Magalhães. Cross-modal subspace learning with scheduled adaptive margin constraints. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 75–83, 2019. 3

[41] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding. In *IJCAI*, 2018. 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[43] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xiansheng Hua. Disentangled representation learning for text-video retrieval. *ArXiv*, abs/2203.07111, 2022. 2

[44] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14020–14029, 2021. 3

[45] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 2

[46] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1459–1468, 2021. 3

[47] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 5

[48] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6389–6394, 2019. 5

[49] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713, 2023. 2, 3

[50] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1645–1653. Association for Computing Machinery, 2017. 6

[51] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[52] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 2, 3

[53] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. 5

[54] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 5

[55] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020. 1

[56] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3517–3525, 2022. 3

[57] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 3

[58] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021. 2, 3