Semantic Fusion Augmentation and Semantic Boundary Detection: A Novel Approach to Multi-Target Video Moment Retrieval

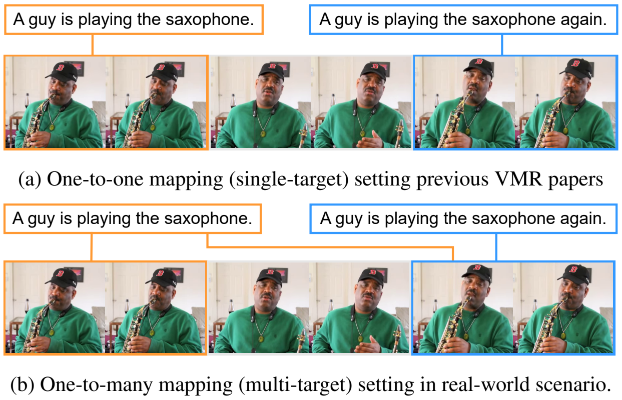
Cheng Huang, Yi-Lun Wu, Hong-Han Shuai, Ching-Chun Huang
National Yang Ming Chiao Tung University, Taiwan
{vin30731.ee10, yilun.ee08, hhshuai, chingchun}@nycu.edu.tw

Abstract

*Given an untrimmed video and a natural language query, video moment retrieval (VMR) aims to retrieve video moments described by the query. However, most existing VMR methods assume a one-to-one mapping between the input query and the target video moment (single-target VMR), disregarding the possibility that a video may contain multiple target moments that match the query description (multi-target VMR). Previous methods tackle multi-target VMR by incorporating false negative moments with the original target moment for multi-target training. However, existing methods cannot properly work when no false negative moments exist in the video, or when the identified false negative moments are noisy but are still being utilized as pseudo-labels. In this paper, we propose to tackle multi-target VMR by **S**emantic **F**usion **A**ugmentation and Semantic **B**oundary **D**etection (**SFABD**). Specifically, we use feature-level augmentation to generate augmented target moments, along with an intra-video contrastive loss to ensure feature consistency. Meanwhile, we perform semantic boundary detection to adaptively remove all false negatives from the negative set of contrastive loss to avoid semantic confusion. Extensive experiments conducted on Charades-STA, ActivityNet Captions, and QVHighlights show that our method achieves state-of-the-art performance on multi-target metrics and single-target metrics. The source code is available at https://github.com/basiclab/SFABD.*

1. Introduction

Recognizing and locating meaningful events within videos is a crucial challenge in computer vision since it demands the model ability of comprehensive understanding. Although there has been notable progress in temporal action localization [21, 24, 25], these models can only identify predefined simple action classes, such as swimming or running. To overcome this limitation, video moment retrieval with natural language query (VMR) has emerged as a more flexible approach to temporal action localization. VMR necessitates models to recognize and localize events described by complex natural language queries, rather than being limited to a predefined set of action classes. As such, VMR can facilitate a variety of real-world applications, such as finding user-defined moments of interest in surveillance videos [9].


A guy is playing the saxophone.
A guy is playing the saxophone again.
A guy is playing the saxophone.
A guy is playing the saxophone again.

(a) One-to-one mapping (single-target) setting previous VMR papers

(b) One-to-many mapping (multi-target) setting in real-world scenario.

Figure 1. The comparison of single-target and multi-target VMR. The query and video are taken from ActivityNet Captions.

Most of the existing Video Moment Retrieval (VMR) methods operate under the single-target assumption, which posits a one-to-one mapping between the input query and its target moment in a video (Figure 1a). This assumption results in some methods predicting only a singular output moment for each input query [6, 12, 16, 17, 19, 28]. The primary rationale behind this assumption is the nature of commonly used VMR datasets, which predominantly offer single-target samples. However, a recent study by [14] highlights the presence of false negative moments in VMR datasets. This study recollected five annotations for some testing samples in Charades-STA [5] and ActivityNet Captions [10], allowing researchers to evaluate the multi-target performance of their models, which are solely trained on single-target samples.

However, training solely on single-target samples can introduce a single-target prediction bias, where models tend to identify only the moment most closely related to the query [34]. Addressing this, [32, 34] pioneered the multi-target assumption to mitigate such biases. Their core concept is to perform multi-target training using only the original single-target annotations. They introduce an auxiliary model branch designed to detect false negative moments in videos that are not explicitly labeled. By incorporating these detected moments with the original single-target labels, multi-target samples are created for multi-target training. This innovative approach has demonstrated significant performance improvements on Charades-STA and ActivityNet Captions datasets in multi-target testing scenarios. However, a limitation of their methods is the reliance on the existence of false negative moments in videos, which are then used as pseudo-labels to create multi-target samples. Additionally, these pseudo-labels can sometimes be noisy, potentially compromising model performance.

To this end, we propose a general framework called **S**emantic **F**usion **A**ugmentation and Semantic **B**oundary **D**etection (**SFABD**) to better tackle multi-target VMR. Specifically, to generate multi-target samples, we propose to utilize feature-level mixup augmentation, which mixes the original positive moment and another randomly selected moment in the video. Then, we propose using intra-video contrastive loss to utilize the property of multi-target labels, ensuring the feature consistency between different positive moments. Given that false negative samples can lead to semantic confusion during optimization [3, 8], we modify the approach from [3] and design an adaptive strategy to determine the semantic boundary to find false negative samples in VMR datasets, which are then removed from the negative set of contrastive loss. Compared to previous multi-target VMR methods [32, 34], our SFABD does not require the existence of false negative moments in the video to generate multi-target samples. Furthermore, the quality of the generated multi-target samples can be maintained by controlling the intensity of the augmentation. Extensive experiments conducted on Charades-STA, ActivityNet Captions, and real multi-target dataset QVHighlights [11] show that our SFABD achieves superior performance compared to previous VMR methods in both multi-target and single-target metrics. The main contributions of this work are summarized as follows.

- We introduce a novel strategy for generating multi-target samples that eliminates the necessity for false negative moments.
- We propose an intra-video contrastive loss for multi-target samples, along with an adaptive strategy for false negative elimination.
- Experimental results demonstrate that our proposed SFABD achieves state-of-the-art performance on both multi-target and single-target datasets.

2. Related Works

Single-Target VMR. Various methods have been proposed to address the problem of VMR, and they can be broadly categorized into proposal-free and proposal-based methods. Proposal-free methods treat VMR as either a regression or classification problem. Some approaches [6, 12, 16, 27, 29, 33] directly regress the start and end timestamps of the target moment. Others classify the probability that each frame is the starting or ending frame [6, 17, 19, 28]. However, a major limitation of most proposal-free methods is that they are designed to predict a single output moment. While these methods perform well on simgle-target samples, they are unable to handle multi-target samples due to their inability to generate multiple predictions. On the other hand, proposal-based methods tackle VMR using a proposal-and-rank framework. Approaches such as [1, 5] employ sliding windows to generate multi-scale proposals, which are then compared independently with the query to obtain similarity scores. [2] proposed to use a predefined set of anchors starting from each frame to efficiently compute similarity scores for multi-scale proposals. More recently, [31] introduced temporal max-pooling to construct a 2D temporal proposal map and employed convolutional networks to capture relationships between neighboring proposals. Proposal-based methods have a natural advantage in generalizing to multi-target VMR compared to most proposal-free methods, primarily due to their ability to generate multiple predictions.

Multi-Target VMR. While VMR methods with multiple predictions can be generalized to multi-target testing scenarios, their optimization goals still primarily focus on single-target prediction. Models trained solely on single-target samples tend to develop a bias towards predicting the most matched moment with the query, rather than generating diverse predictions that encompass all target moments [34]. To address this issue, [34] leveraged the verbs and nouns in the query to identify false negative moments within the same video, which were then used as pseudo-labels for multi-target training. Similarly, [32] proposed the use of video captioning techniques to reconstruct the query for each proposal moment, allowing them to measure the similarity scores between the reconstructed queries and the input query to identify false negative moments for multi-target training. However, both [34] and [32] rely on the existence of false negative moments in the video to generate multi-target samples. Moreover, the quality of the generate multi-target samples cannot be guaranteed.

False Negative Detection. False negative samples share the same semantic concept as the query, but are mistakenly assigned to the negative set of contrastive loss. This may lead to the model discarding shared semantic information between the positive sample and the false negative sample [8]. To perform false negative detection, approaches such as [32, 34] consider negative samples whose similar-

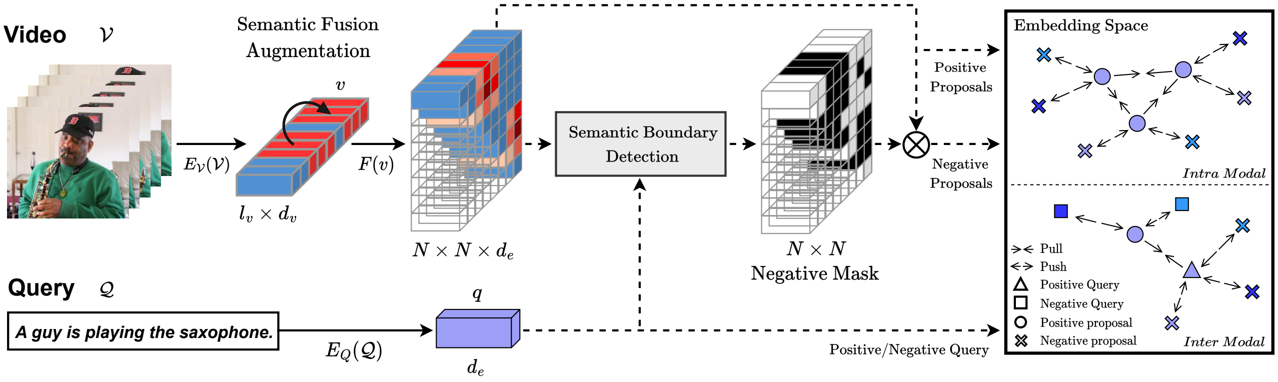


Figure 2. The pipeline of our proposed SFABD. Based on the video feature extracted by the pre-trained video encoder, we perform semantic fusion augmentation for generating augmented target moments. Afterward, we transform the sequence into a 2D temporal proposal map, following by a proposal encoder. The semantic boundary detection is then performed to adaptively detect false negative samples and remove them by generating a proper negative mask. Finally, the video proposals and queries are projected into a common embedding space for contrastive learning. (The video proposals and queries from different batch indices are denoted in different colors in this figure.)

ity score with the anchor (query in VMR) exceeds a fixed threshold as false negative samples. [3] mentioned that false negative samples found at the early stage of training are not reliable and will gradually become more reliable as training progresses. Therefore, they design a candidate-and-accept method that first finds false negative candidates by clustering, and then gradually accepts the candidates by linearly increasing the acceptance rate. Once false negative samples are identified, there are two common approaches to handle them. The first is false negative attraction, which uses false negative samples as pseudo-labels [8, 32, 34]. The second is false negative elimination, which removes false negative samples from the negative set [3]. Although false negative attraction may improve performance by increasing data diversity, it could also degrade performance if the found false negative samples are noisy. On the other hand, false negative elimination is more tolerant to noisy false negative samples, making it a more stable option.

3. Method

3.1. Problem Formulation

Given an untrimmed video $\mathcal{V}$ and a natural language query $\mathcal{Q}$, VMR aims to predict temporal moments $\mathcal{M} \in \mathbb{R}^{K\times 2}$ that semantically match the input query, where K is the number of predicted moments and the second dimension represents the start and end timestamps of the predicted moment. The ground truth moments are denoted as $\hat{\mathcal{M}} \in \mathbb{R}^{\hat{K}\times 2}$, where $\hat{K}$ is the number of ground truth moments. It is important to note that the terms “target moments”, “positive moments”, and “ground truth moments” are interchangeable in this context.

We use $v = E_{\mathcal{V}}(\mathcal{V}) \in \mathbb{R}^{l_v \times d_v}$ to denote the video features obtained by a pre-trained video encoder $E_{\mathcal{V}}$, where l_v is the length of the extracted features and d_v is the dimension of feature space. As for the query, we use $q = E_Q(\mathcal{Q}) \in \mathbb{R}^{d_e}$ to represent the query feature, where E_Q is the pre-trained sentence encoder and d_e is the dimension of the query feature space.

3.2. Method Overview

To enable multi-target training, we first generate multi-target samples through semantic fusion augmentation (Sec.3.3). Subsequently, we transform the video features into a 2D proposal map using the same procedure as [22, 31]. Finally, we utilize semantic boundary detection to adaptively detect false negative samples (Sec.3.5) and remove them from the negative set of both contrastive loss (Sec.3.4) and cross-entropy loss (Sec.3.6).

3.3. Semantic Fusion Augmentation (SFA)

To enable multi-target training on datasets that contain only single-target samples, we employ mixup augmentation [4, 20, 23, 26] on the video feature to generate augmented target moments. Specifically, we blend the feature sequence of the target moment into the feature sequence of a randomly selected background moment using a weighted averaging approach. This process expands the number of target moments in $\mathcal{V}$ without requiring direct access to the original video. We have observed several advantages of employing feature-level augmentations instead of frame-level augmentations. First, the video encoding process is not required. When using VGG as the video encoder in Charades-STA, a training speed of approximately 10 times faster can be achieved using feature-level augmentation compared to frame-level augmentation. Most importantly, as shown later in ablation study (Sec.4.4), the performance gap is negligible. Second, obtaining the data pre-processing details and the pre-trained weight of each video encoder is not required. In VMR, the data-preprocessing details and the pre-trained weight of the video encoder are usually not provided.

3.4. Intra-Modal Contrastive Loss

When conducting multi-target training, it is advantageous to leverage the inherent label structure within multi-target samples. Specifically, all positive moments within the multi-target samples share the same semantic of the query; hence the proposal encoder should extract consistent proposal features from these positive moments. To ensure and guide this behavior, we propose using intra-video contrastive loss specifically designed for multi-target samples. We consider a batch of video features $\{v_b\}_{b=1}^{B}$, along with their corresponding proposal features $\{F(v_b)\}_{b=1}^{B}$, where $F\colon \mathbb{R}^{l_v \times d_v} \to \mathbb{R}^{N \times N \times d_e}$ represents the proposal encoder [22, 31], B denotes the batch size, d_e indicates the dimension of the embedding space, and N is the number of units to which the video feature is divided. To simplify the notation, we use a one-dimensional index $i \in [1 \mathinner{..} N']$ to refer to the proposal feature at position $(x, y) \in [1 \mathinner{..} N]^2$ within a two-dimensional feature matrix, denoted as $f_i^b = F(v_b)_{xy} \in \mathbb{R}^{d_e}$, where the superscript $b \in [1 \mathinner{..} B]$ refers to the index within the batch, N' represents the number of proposals in the 2D feature matrix, x and y indicate the start time and the end time, respectively. For a query $\mathcal{Q}^b$, we calculate the intersection over union (IoU) between the proposal moments and the target moments. We extract the top-k proposals with the highest IoU for each target moment as positive proposals. The indices of these positive proposals are represented as a set I_+^b.

For the positive proposal pair $(f_i^b, f_{i_+}^b)$ where $i, i_+ \in I_+^b$, the proposed intra-modal contrastive loss is formulated as follows:

$$p(f_{i_+}^b \mid f_i^b) = \frac{\exp\left((s_{ii_+}^{bb} - m_{\mathrm{vv}})/\tau_{\mathrm{vv}}\right)}{\exp\left((s_{ii_+}^{bb} - m_{\mathrm{vv}})/\tau_{\mathrm{vv}}\right) + \sum_{b'=1}^{B}\sum_{j=1}^{N'} M_j^{b'} \exp\left(s_{ij}^{bb'}/\tau_{\mathrm{vv}}\right)}, \tag{1}$$

where $s_{ij}^{ab} = f_i^a \cdot f_j^b / (\|f_i^a\| \cdot \|f_j^b\|)$ represents the cosine similarity between f_i^a and f_j^b, m_{vv} is the intra-video margin, and τ_{vv} is the intra-video temperature. Furthermore, $M_j^{b'}$ represents the negative mask that indicates which proposal feature $f_j^{b'}$ is negative *w.r.t.* f_i^b. Specifically,

$$M_j^{b'} = \begin{cases} 0 & \text{if } b' = b \text{ and } j \in I_+^b \\ 1 & \text{Otherwise} \end{cases}. \tag{2}$$

For consistency, we re-express the inter-modal contrastive loss [22] of our backbone using our notation:

$$p(q^b \mid f_{i_+}^b) = \frac{\exp\left((\phi_{i_+}^{bb} - m_{\mathrm{vq}})/\tau_{\mathrm{vq}}\right)}{\exp\left((\phi_{i_+}^{bb} - m_{\mathrm{vq}})/\tau_{\mathrm{vq}}\right) + \sum_{b'=1, b' \neq b}^{B} \exp\left(\phi_{i_+}^{bb'}/\tau_{\mathrm{vq}}\right)}, \tag{3}$$

$$p(f_{i_+}^b \mid q^b) = \frac{\exp\left((\phi_{i_+}^{bb} - m_{\mathrm{qv}})/\tau_{\mathrm{qv}}\right)}{\exp\left((\phi_{i_+}^{bb} - m_{\mathrm{qv}})/\tau_{\mathrm{qv}}\right) + \sum_{b'=1}^{B}\sum_{j=1}^{N'} M_j^{b'} \exp\left(\phi_j^{b'b}/\tau_{\mathrm{qv}}\right)}. \tag{4}$$

Here $\phi_i^{ab} = f_i^a \cdot q^b / (\|f_i^a\| \cdot \|q^b\|)$ represents the cosine similarity between the proposal feature f_i^a and the query feature $q^b = E_q(\mathcal{Q}^b)$, where a and b refer to the indices within the batch. Additionally, m_{vq} and τ_{vq} denote the inter-video margin and temperature, respectively. Similarly, m_{qv} and τ_{qv} denote the inter-query margin and temperature, respectively. The negative mask $M_j^{b'}$ is defined in Eq.(2).

3.5. Semantic Boundary Detection (SBD)

To accurately identify false negative samples, setting a proper semantic boundary is crucial. However, previous approaches [3,32,34] simply set a fixed semantic boundary for all samples to detect false negative samples, without considering the differences between each sample. We modify the candidate-and-accept approach in [3] and design an adaptive strategy to determine the proper semantic boundary for each sample, which takes the differences between each sample into consideration.

Adaptive Threshold. Instead of using clustering to find false negative candidates [3], we adopt a more efficient threshold-based method. We assume that the query is at the center of its semantic cluster, allowing us to build a set of false negative candidates by comparing the similarities between the negative samples and the query. Instead of using a fixed similarity threshold for all queries [32, 34], we propose using the maximum similarity between query and positive proposals to identify false negative candidates.

$$\gamma^b = \max_{i_+} \{\phi_{i_+}^{bb}\}. \tag{5}$$

This approach takes into account the varying distributions of semantic clusters in the embedding space, as illustrated in Figure 3. By setting the threshold based on the maximum positive similarity, we can adapt to the sparsity or density of the query semantic cluster. This ensures that false negatives are not omitted in sparse clusters and avoids misclassifying hard negatives as false negatives in dense clusters.

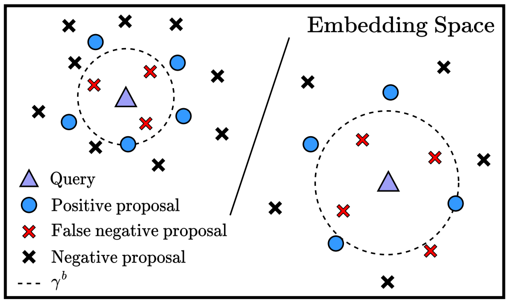


Figure 3. Illustration of a dense cluster (upper left) and a sparse cluster (lower right) for adaptive threshold. Best viewed in color.

Adaptive Acceptance Rate. [3] assumes that all samples will be progressively learned by the model, as depicted in Figure 4a. Therefore, they use the same linearly increasing acceptance rate to gradually trust the false negative candidates for all samples. However, we observed that not all samples exhibit learning progress that closely matches the ideal scenario. As shown in Figure 4b, the curve μ_+^b does not show a consistent improvement and is sometimes lower than that of μ_-^b. Importantly, when the contrastive gap $\delta^b = \mu_+^b - \mu_-^b$ is lower or even negative, it indicates that the sample is not well represented in the embedding space. Therefore, the use of the same linearly scheduled acceptance rate for all samples is not a favorable choice. To address the variable learning progress of each sample, we propose using the contrastive gap δ^b as an indicator. A larger contrastive gap implies that positive and negative samples can be easily separated by the learned representation, while a smaller contrastive gap suggests that the model struggles to distinguish between positive and negative proposals. We employ a simple function to transform the contrastive gap into the acceptance rate:

$$r^b = \max(\min(\alpha \cdot \delta^b, 1), 0), \tag{6}$$

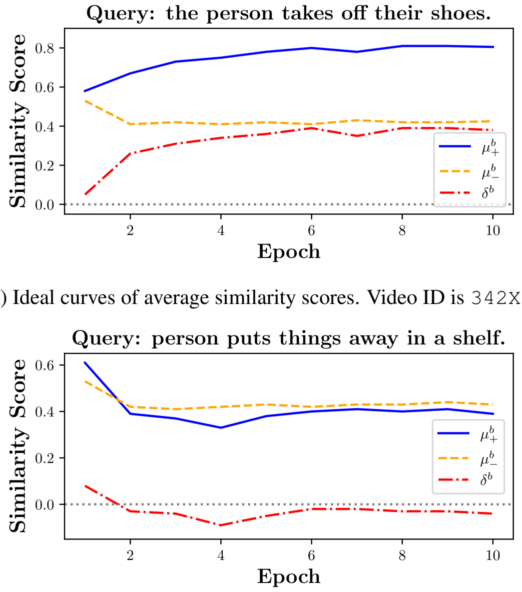
where $\alpha \in \mathbb{R}^+$ is a constant hyper-parameter.

The procedure of our SBD is summarized as follows. First, we use the adaptive threshold γ^b to identify false negative candidates for the query feature q^b:

$$\Phi^b = \left\{ \phi_i^{bb} \,\middle|\, i \in [1\,..\,N'], i \notin I_+^b, \phi_i^{bb} \geq \gamma^b \right\}, \tag{7}$$

where I_+^b represents the positive proposals defined in Sec. 3.4. The next step involves selecting the top-k elements from Φ^b, based on the adaptive acceptance rate r^b:

$$I^b_{\text{false}} = \left\{ i \,\middle|\, \phi_i^{bb} \text{ is in top } k \text{ of } \Phi^b \right\}, k = \left\lceil r^b \cdot \left|\Phi^b\right| \right\rceil. \tag{8}$$



(a) Ideal curves of average similarity scores. Video ID is `342XO`.

(b) Unfavorable curves of average similarity scores. Video ID is `8YKGP`.

Figure 4. Average similarity scores of ideal and unfavorable samples in the Charades-STA dataset. Here, $\mu_+^b = \sum_{i_+} \phi_{i_+}^{bb} / \hat{K}$ represents the average positive pair similarity scores for query feature q^b, and $\mu_-^b = \sum_{b'=1}^{B} \sum_{i=1}^{N'} \phi_i^{b'b} / (BN' - K)$ represents the average negative pair similarity scores for q^b, where b refers to the index within the batch, $\hat{K}$ is the number of ground truth moments and $\delta^b = \mu_+^b - \mu_-^b$. Best viewed in color.

For Eq.(1) and Eq.(4), we utilize the following negative mask when SBD is applied:

$$M_j^{b'} = \begin{cases} 0 & \text{if } b' = b \text{ and } j \in I_+^b \cup I_{\text{false}}^b \\ 1 & \text{Otherwise} \end{cases}. \tag{9}$$

3.6. Loss Functions

Similar to previous approaches [22, 31], we adopt the scaled IoU loss $\mathcal{L}_{\text{iou}}$. It is important to note that the false negative proposals in I_{false}^b are also eliminated from the scaled IoU loss. Additionally, the contrastive constraints are learned by optimizing the negative log likelihood:

$$\mathcal{L}_{\text{intra}} = \frac{-1}{B|I_+^b|^2} \sum_{b=1}^{B} \sum_{i, i_+ \in I_+^b} \log p(f_{i_+}^b \mid f_i^b) \tag{10}$$

$$\mathcal{L}_{\text{inter}} = \frac{-1}{B|I_+^b|} \sum_{b=1}^{B} \sum_{i_+ \in I_+^b} \left(\log p(q^b \mid f_{i_+}^b) + \log p(f_{i_+}^b \mid q^b) \right) \tag{11}$$

The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{iou}} + \lambda_{\text{inter}} \mathcal{L}_{\text{inter}} + \lambda_{\text{intra}} \mathcal{L}_{\text{intra}}, \tag{12}$$

where λ_{inter} and λ_{intra} are loss weights.

4. Experiments

4.1. Datasets

Charades-STA. (Charades) [5] This dataset consists of 5,338 videos with 12,408 single-target queries in the training set and 1,334 videos with 3,720 single-target queries in the testing set. Moreover, [14] proposed a relabeled test set with 1,000 queries and 5 target moments for each query, and we refer to this testing set as Charades-multi.

ActivityNet Captions. (ActivityNet) [10] This dataset consists of 10,009 videos with 37,417/17,505/17,031 single-target queries in the training/validation/testing set. Additionally, [14] proposed a relabeled test set with 1,288 queries and 5 target moments for each query, and we refer to this testing set as ActivityNet-multi.

QVHighlights. [11] This dataset is a high-quality multi-target dataset that consists of 10,310 multi-target queries associated with 18,367 moments in 10,148 videos. The videos are recently collected from YouTube and contain content from three main categories: Daily Vlog, Travel Vlog, and News. This dataset serves as a fair benchmark as the testing performance can only be evaluated by submitting predictions to the official evaluation server[1].

4.2. Evaluation Metrics

Following previous work on single-target VMR, we use the R@n, IoU=m metric [5] to evaluate single-target datasets. This metric measures the percentage of queries that have at least one correctly retrieved moment (IoU $> m$) among the top-n output moments. To evaluate Charades-multi and ActivityNet-multi, we adopt the R@(n, G), IoU=m metric [33]. This metric calculates the percentage of the G target moments that have at least one matched prediction (IoU $> m$) among the top-n predicted moments. To evaluate QVHighlights, we use the mean average precision mAP@m metric [11], where m represents the IoU threshold for correct detection. Average results across IoU thresholds ranging from 0.5 to 0.95 (inclusive) with a step size of 0.05 are denoted as mAP@avg.

4.3. Evaluation Results

For the sake of fairness, we only compare methods that use the same pre-trained video encoder. If a different input data source is utilized, it will be clearly specified in the tables. When comparing our method with existing approaches on Charades-multi and ActivityNet-multi, as shown in Table 1 and Table 2, our method outperforms all previous methods. Unlike DTG and DTG-SPL that rely on the existence of false negative moments in a video, our method does not have such a limitation. Even if the video does not contain any false negative moments, our method can still generate high-quality multi-target samples, resulting in consistently superior performance.

Method	Feature	R@(5,5)	
		IoU=0.5	IoU=0.7
2DTAN [31]	VGG	49.30	-
MMN† [22]	VGG	56.68	30.16
SFABD (Ours)	VGG	**57.62**	**30.52**
DRN [27]	C3D	46.63	-
SFABD (Ours)	C3D	**58.38**	**31.06**
2DTAN [31]	I3D	54.56	-
DeNet [33]	I3D	56.30	-
DTG [34]	I3D	60.72	-
DTG-SPL [32]	I3D	61.88	-
SFABD (Ours)	I3D	**65.14**	**36.16**

Table 1. Evaluation results on Charades-multi. Note that † means that the results are evaluated on the official pre-trained model.

Method	Feature	R@(5,5)	
		IoU=0.5	IoU=0.7
2DTAN [31]	C3D	56.35	-
DeNet [33]	C3D	58.46	-
MMN† [22]	C3D	59.39	42.47
SFABD (Ours)	C3D	**61.97**	**44.80**
2DTAN [31]	I3D	55.08	-
DTG [34]	I3D	58.51	-
DTG-SPL [32]	I3D	59.32	-
SFABD (Ours)	I3D	**60.79**	**43.77**

Table 2. Evaluation results on ActivityNet-multi. Note that † means that the results are evaluated on the official pre-trained model.

Method	mAP@0.5	mAP@0.75	mAP@avg
momentDETR [11]	60.51	35.36	36.14
UMT† [13]	53.38	37.01	38.08
QD-DETR [15]	62.52	39.88	39.86
QD-DETR† [15]	**63.04**	40.10	40.19
SFABD (Ours)	62.38	**44.39**	**43.79**

Table 3. Evaluation results on the QVHighlights test set. The symbol † indicates that the source feature includes video and audio. Otherwise, the input feature consists of video only.

Table 3 shows the results on QVHighlights. Our method achieves current state-of-the-art and has a large performance gap compared to previous work [15]. The major difference between our method and previous methods lies in the intra-video modality. Previous methods all adopted transformer-based model that used self-attention for video sequence encoding. However, their self-attention process was unsupervised and did not fully utilize the information in multi-target labels. In contrast, our method fully utilizes the label information by using the intra-video contrastive loss to supervise the relationships between positive moments. This aspect is particularly important in multi-target VMR where different positive moments may share the same semantics but have a very different visual appearance.

In addition, we also evaluate our method in the single-

[1] https://codalab.lisn.upsaclay.fr/competitions/6937

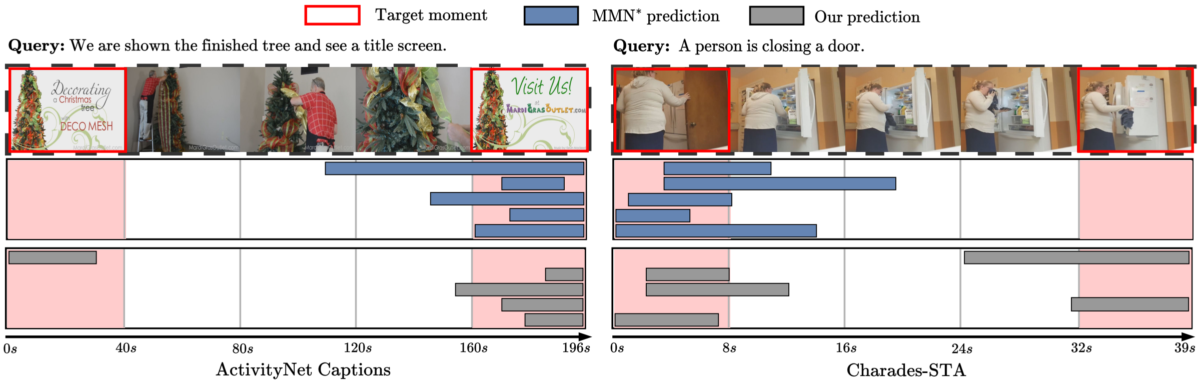


Figure 5. Visualization of multi-target prediction. It is worth noting that MMN* is implemented in our code base.

Method	C3D video features		I3D video features	
	R@1	R@5	R@1	R@5
	IoU=0.7	IoU=0.7	IoU=0.7	IoU=0.7
DRN [27]	26.40	55.38	31.75	60.05
VSLNet [28]	<u>30.19</u>	-	-	-
MS-2D-TAN [30]	23.25	48.55	36.21	61.13
DTG [34]	-	-	39.38	66.91
DTG-SPL [32]	-	-	40.13	67.12
BMRN [18]	28.37	<u>57.19</u>	**42.46**	<u>67.65</u>
SFABD (Ours)	**30.51**	**59.96**	<u>40.21</u>	**68.65**

Table 4. Evaluation results on Charades with the C3D feature and the I3D feature. Due to space limitations, the results for IoU=0.5 are provided in Appendix E.

Method	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
VSLNet [28]	43.22	26.16	-	-
MS-2D-TAN [30]	45.50	<u>28.28</u>	<u>79.36</u>	<u>61.70</u>
DTG [34]	47.03	26.12	78.47	59.54
DTG-SPL [32]	<u>47.04</u>	-	79.16	-
SFABD (Ours)	**49.22**	**30.97**	**81.03**	**66.81**

Table 5. Evaluation results on ActivityNet with the I3D feature.

target scenario to ensure that the single-target performance is maintained. It turns out that our method establishes new state-of-the-arts in some settings. Table 4 shows that our method achieves comparable results compared to the recent single-target VMR method [18]. Furthermore, Table 5 shows that our method outperforms recent multi-target VMR methods [32, 34] and previous single-target VMR methods on ActivityNet using the I3D feature. The incorporation of SFA and intra-video contrastive loss actually shares a similar concept with self-supervised image pretraining, which ensures the image encoder to extract consistent features of the original image and the augmented one, resulting in a more robust image encoder. Therefore, a more robust and consistent proposal encoder is learned by using SFA and intra-video contrastive loss. Our semantic boundary detection further eliminates the negative effect of false negative samples during batch-wise contrastive learning, resulting in further improvement.

We would like to clarify the potential challenge of achieving a concurrent improvement between R@(5,5) and R@1 in Charades and ActivityNet. The core issue stems from the assumption that target moments in a multi-target sample hold equal priority. Consequently, even a well-trained multi-target model cannot ensure that the prediction with the highest confidence aligns precisely with a single-target label. As a result, R@(5,5) cannot guarantee simultaneous enhancement with R@1. For illustrative examples of this phenomenon, please consult Appendix D. We consider this to be a primary factor that contributes to the slightly lagging performance of SFABD in Charades I3D features compared to BMRN [18]. However, it is worth noting that while BMRN concentrates on refining the proposal boundaries of the 2D temporal proposal map, our approach focuses on enhancing proposal features for multi-target VMR. These two methods are orthogonal and can be synergistically combined to create a powerful model.

4.4. Ablation Studies

Method Ablation. We further investigate the effectiveness of each module in our proposed method, including SFA, intra-video contrastive loss, and SBD. In Table 6, we present the ablation results of Charades with the VGG feature, while the results for other datasets can be found in the Appendix E. We observe that SFA (second row) can slightly increase overall performance by adding data diversity. Using SBD (third row) also improves overall performance compared to our baseline model [22] (first row) by eliminating the negative effect of false negative samples. Combining SFA with intra-video contrastive loss (fourth row) leads to further improvements by utilizing the multi-target labels information to ensure feature consistency between positive moments. Using all methods together (fifth

$\mathcal{L}_{\text{inter}}$	SFA	$\mathcal{L}_{\text{intra}}$	SBD	R@1		R@5		R@(5,5)	
				IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
✓				48.24	29.15	83.79	60.01	55.46	30.56
✓	✓			48.11	29.61	84.96	59.00	57.26	30.60
✓			✓	48.82	29.50	85.04	60.17	57.1	**30.66**
✓	✓	✓		48.60	29.96	**85.94**	60.45	57.22	30.36
✓	✓	✓	✓	**50.23**	**31.38**	85.62	**61.07**	**57.62**	30.52

Table 6. Full Method Ablation on Charades-STA with the VGG Feature.

Method	Proposal Encoder	R@1		R@5	
		IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
MMN	ConvNet	47.31	27.28	83.74	58.41
MMN†	ConvNet	46.99	27.95	83.33	59.41
SFABD	ConvNet	47.04	29.50	83.71	59.14
MMN†	ResNet-18	48.24	29.15	83.79	60.01
SFABD	ResNet-18	**50.23**	**31.38**	**85.62**	**61.07**

Table 7. Evaluation results on Charades with the VGG feature and different proposal encoders. ConvNet is the encoder officially used by MMN [22]. Note that † denotes that the implementation is based on our code base.

Threshold	Acceptance Rate	R@1	R@5	R@(5,5)
		IoU=0.7	IoU=0.7	IoU=0.7
Fixed	Linear	29.79	65.36	43.66
Fixed	Adaptive	30.16	65.95	42.61
Adaptive	Linear	30.47	66.14	43.51
Adaptive	Adaptive	**30.97**	**66.81**	**43.77**

Table 8. SBD ablation on ActivityNet with the I3D feature.

row), the best overall performance is achieved.

Proposal Encoder. Table 7 compares the performance of different proposal encoders. To provide a fair comparison, we conducted a hyperparameter search for MMN when combined with ResNet-18 [7]. The results demonstrate that our method outperforms MMN using both types of encoder. By transitioning from a ConvNet to ResNet-18, our approach achieves a more substantial performance improvement compared to MMN. One possible explanation is that ResNet-18 possesses a higher learning capacity than a simple ConvNet, enabling it to recognize more patterns with increased supervision.

Semantic Boundary Detection. In Table 8, we compare the performance of different threshold strategies and different acceptance rate schedulers. For the experimental details, please refer to Appendix C. Using a fixed threshold to find false negative candidates regardless of differences between semantic clusters leads to inferior performance. Similarly, using the same linearly increasing acceptance rate [3] for all samples regardless of their learning progress also leads to inferior performance. The best results are achieved by using an adaptive strategy on both the threshold and the acceptance rate.

Feature-Level Augmentation We compared the effects of data augmentation using mixup and cutmix techniques on both original videos and VGG features. For the results and experimental details, please refer to Appendix C.

4.5. Visualizations

Figure 5 visualizes the predictions of MMN [22] and our SFABD . MMN is trained only with single-target samples, therefore all of their predictions tend to focus on the most related target moment to the query. In contrast, our SFABD successfully predicts the other target moment due to the multi-target training enabled by Semantic Fusion Augmentation. The visualization for the false negative moments found by Semantic Boundary Detection is in Appendix A.

5. Conclusion

In this paper, we propose an efficient and concise method SFABD to tackle multi-target video moment retrieval. Specifically, we use Semantic Fusion Augmentation to generate multi-target samples for multi-target training, and utilize an intra-video contrastive loss to ensure feature consistency among different positive moments. Furthermore, we employ Semantic Boundary Detection to adaptively eliminate false negative moments from the negative set of contrastive loss. Our SFABD does not rely on the existence of false negative moments in videos, and further ensures the quantity and quality of the generated multi-target samples. The extensive experiments show that our SFABD achieves state-of-the-art performance on QVHighlights, Charades-STA and ActivityNet Captions datasets.

6. Future Works

The current SBD pipeline consists of two steps: finding false negative candidates using an adaptive threshold and accepting a portion of them based on an adaptive acceptance rate. In the future, we aim to integrate these two steps into a unified process that can achieve the same objective concurrently.

Acknowledgement

This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC-112-2425-H-A49-001, NSTC-109-2221-E-009-114-MY3, NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5803–5812, 2017. 2

[2] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, 2018. 2

[3] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. In *International Conference on Learning Representations*, 2022. 2, 3, 4, 5, 8

[4] Alex Falcon, Giuseppe Serra, and Oswald Lanz. A feature-space multimodal data augmentation technique for text-video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4385–4394, 2022. 3

[5] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5267–5275, 2017. 1, 2, 6

[6] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[8] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 2785–2795, 2022. 2, 3

[9] Abdolamir Karbalaie, Farhad Abtahi, and Mårten Sjöström. Event detection in surveillance videos: a review. *Multimedia Tools and Applications*, 81(24):35463–35501, 2022. 1

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 706–715, 2017. 1, 6

[11] 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

[12] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1902–1910, 2021. 1, 2

[13] 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. 6

[14] Otani Mayu, Nakahima Yuta, Rahtu Esa, and Heikkilä Janne. Uncovering hidden challenges in query-based video moment retrieval. In *The British Machine Vision Conference (BMVC)*, 2020. 1, 6

[15] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 6

[16] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 1, 2

[17] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 2464–2473, 2020. 1, 2

[18] Muah Seol, Jonghee Kim, and Jinyoung Moon. Bmrn: Boundary matching and refinement network for temporal moment localization with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5570–5578, 2023. 7

[19] Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. Frame-wise cross-modal matching for video moment retrieval. *IEEE Transactions on Multimedia*, 24:1338–1349, 2021. 1, 2

[20] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 3

[21] Limin Wang, Yu Qiao, Xiaoou Tang, et al. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2):2, 2014. 1

[22] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2613–2623, 2022. 3, 4, 5, 6, 7, 8

[23] Han Wu, Chunfeng Song, Shaolong Yue, Zhenyu Wang, Jun Xiao, and Yanyang Liu. Dynamic video mix-up for cross-domain action recognition. *Neurocomputing*, 471:358–368, 2022. 3

[24] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2016. 1

[25] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kassim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016. 1

[26] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020. 3

[27] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2, 6, 7

[28] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4252–4266, 2021. 1, 2, 7

[29] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021. 2

[30] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9073–9087, 2021. 7

[31] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020. 2, 3, 4, 5, 6

[32] Hao Zhou, Chongyang Zhang, Yanjun Chen, and Chuanping Hu. Towards diverse temporal grounding under single positive labels. *arXiv preprint arXiv:2303.06545*, 2023. 2, 3, 4, 6, 7

[33] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8454, 2021. 2, 6

[34] Hao Zhou, Chongyang Zhang, Yan Luo, Chuanping Hu, and Wenjun Zhang. Thinking inside uncertainty: Interest moment perception for diverse temporal grounding. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7190–7203, 2022. 2, 3, 4, 6, 7