

Our supplementary material covers the following topics: 1) limitations of our algorithm, 2) potential negative societal impact of our algorithm, 3) more details of datasets and implementation, 4) detailed explanation for selection strategies and rule-based negative proposals, 5) experimental results for analysis on dataset biases, reproduced existing methods, and loss balances, and 6) qualitative comparisons to verify the effectiveness of our negative proposals.

A LIMITATIONS

We observe that some negative proposals overlap with positive proposals at the late training stage. This means that the positive proposals are not sufficiently distinguished from the negative proposals. Therefore, further research on contrastive learning for more efficient discrimination could be future work.

B POTENTIAL NEGATIVE SOCIETAL IMPACT

Our algorithm can automatically localize video segments relevant to a given natural language. This ability raises concerns about its potential misuse by an individual or group for malevolent purposes. For example, our algorithm could be used for automated censorship of videos, thereby limiting the freedom of expression and open discourse. To address this concern, we make our algorithm only available for research purposes.

C DATASETS

Charades-STA For video moment localization tasks, Charades-STA dataset [13], which originated from [35], provides videos of daily indoor activities. In the videos, 7,986 videos are used for a training set and 1,863 videos are used for a testing set. On average, the videos are 30 seconds long and have 2.4 annotated temporal boundaries corresponding to the sentence queries. The size of the vocabulary is 1,111. Among 16,128 pairs of a video and a sentence query, there are 12,408 training data and 3,720 testing data. Charades-STA dataset includes data from human subjects and we are not sure if this dataset is approved by an **Institutional Review Board**. Nevertheless, this dataset is still available and has not been withdrawn. Moreover, this dataset is extensively utilized in various applications, including but not limited to action recognition, video captioning, and video moment localization. In order to ensure fair comparisons with other weakly supervised moment localization methods (e.g., [6, 15, 28, 51, 59]), it is imperative to conduct evaluations using this dataset.

ActivityNet Captions For video captioning and video moment localization tasks, ActivityNet Captions dataset [23] provides 20k untrimmed videos from YouTube with 100k sentence queries. In the videos, 10,009 videos are used for a training set, 4,917 videos are used for two validation sets (val_1 and val_2), and 5,044 videos are used for a testing set. On average, the videos are 120 seconds long and have 3.65 annotated temporal boundaries corresponding to the sentence queries. The size of the vocabulary is 8,000. Among 71,953 pairs of a video and a sentence query, there are 37,417 training data, 17,505 validating data (val_1), and 17,031 validating data (val_2). Following [55], we use val_2 as a testing set. ActivityNet Captions dataset includes data from human subjects and we have confirmed that this dataset is approved by the Stanford **Institutional Review Board**.

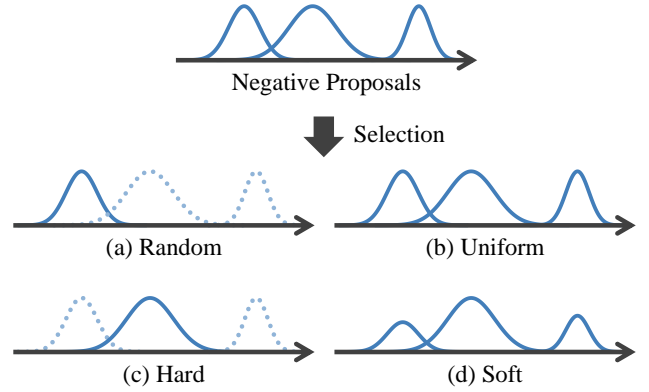


Figure 8: Selection strategies.

TV show Retrieval For a video moment localization task, TV show Retrieval (TVR) dataset [25] provides 21.8k videos from 6 TV shows of diverse genres, with 109k sentence queries. Among 109K pairs of a video and a sentence query, 87.2K pairs are used for a training set, 10.9K pairs are used for a validation set, and 10.9K pairs are used for a testing set. Since the testing set is preserved for the challenge, we evaluate our method on the validation set. Each video is 76.2 seconds long on average. The size of the vocabulary is 57,100. TV show Retrieval dataset includes data from human subjects and we have confirmed that this dataset is approved by the **Institutional Review Board**.

D IMPLEMENTATION DETAILS

For both training and inference, the used GPU and OS are GeForce RTX 4090 and Ubuntu 18.04.06, respectively. Following previous methods for fair comparisons, we pre-train the C3D on sport1M [19] and I3D on Kinetics [7], respectively.

E DETAILS FOR SELECTION STRATEGIES

In the selection network, we use four selection strategies to select useful proposals for query prediction among multiple negative proposals. The four selection strategies are depicted in Fig. 8. Our selection strategies are as follows: ‘None’: select none (no negative proposal is used), ‘Random’: randomly select one, ‘Uniform’: select all with the same selection weights, ‘Hard’: select one with the highest learnable selection weight, and ‘Soft’: select all with different learnable selection weights. Among them, the soft selection strategy and the hard selection strategy have learnable selection weights. Details of the soft selection strategy are described in Sec. 3.2. For the hard selection strategy, using the last outputs from the decoder of the mask-conditioned transformer, we can estimate M selection weights for M negative proposals by two fully connected layers followed by Gumbel Softmax [16]. The selection weights can be written as $[s_1, s_2, \dots, s_M]$, where s_m is 0 or 1 for all m .

Table 6: Analysis on dataset biases and comparisons with fully supervised methods at R@1,IoU=0.5 on Charades-STA and at R@1,IoU=0.3 on ActivityNet Captions. CD-iid and CD-ood are test sets of re-splitted datasets in [52].

Method		Charades-STA			ActivityNet Captions		
		original	CD-iid	CD-ood	original	CD-iid	CD-ood
Fully Supervised	DRN [54]	53.09	41.91	30.43	-	48.92	36.86
	SCDM [53]	54.44	47.36	41.60	54.80	46.44	31.56
Weakly Supervised	WS-DEC [11]	-	14.06	23.67	41.98	26.06	17.00
	Ours	52.47	46.83	42.94	59.12	52.86	40.33

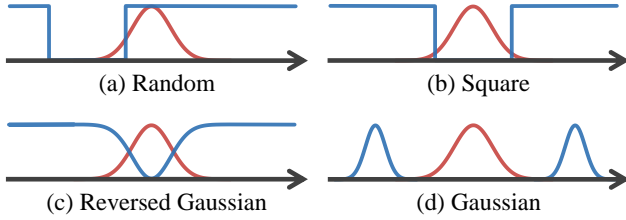


Figure 9: Rule-based negative proposals.

F DETAILS FOR RULE-BASED NEGATIVE PROPOSALS

We use rule-based negative proposals for comparisons with our learning-based negative proposals. In this section, we describe details of the rule-based negative proposals, which are depicted in Fig. 9. We define rule-based negative proposals as follows. ‘Random’: a proposal having a value of zero at the location of a randomly chosen area and a value of one otherwise, ‘Rule-based square’: a proposal having a value of zero at the location of a positive proposal and a value of one otherwise, ‘Rule-based Gaussian’: two proposals of Gaussians whose location is predefined outside both sides of a positive Gaussian proposal, ‘Rule-based reversed Gaussian’: a proposal of Gaussian that is reversed upside down by subtracting a positive Gaussian proposal from a value of one, which is proposed in [58], and ‘Rule-based variable-sized Gaussian’: ‘Rule-based Gaussian’ whose size and location are controlled slightly by the current training epoch, which is proposed in CPL [59].

Even though ‘Rule-based variable-sized Gaussian’ uses the curriculum concept, the negative proposals are not learnable and are always defined to be outside of the positive proposal. Therefore, these negative proposals do not have the ability to capture various confusing locations, because confusing locations also exist inside poorly-generated positive proposals. In contrast, we leverage learnable negative proposals, which are trained by a dual-signed cross-entropy loss with a changing cross-entropy weight, to capture various confusing locations. As the cross-entropy weight changes from minus to plus ones, our negative proposals are gradually changed from easy to hard ones. Tab. 4 shows that our learning-based negative proposal performs much better than the rule-based ones.

G ANALYSIS ON DATASET BIASES

Yuan *et al.* [52] find that both ActivityNet-Captions and Charades-STA datasets have notable moment annotation biases within videos.

In this section, we analyze these dataset biases using re-splitted datasets (*i.e.*, ActivityNet-CD and Charades-CD datasets) proposed in [52]. As shown in Tab. 6, our method makes competitive or higher results than the weakly supervised method [11] and fully supervised methods [53, 54] on test sets (CD-iid and CD-ood) of the re-splitted datasets. It means that our method is less affected by dataset biases. The reason is that our method trains positive and negative proposals by predicting the original sentence queries rather than predicting the biased annotation of temporal locations. Also, our method makes comparable results with fully supervised methods on original datasets.

H REPRODUCED RESULTS OF EXISTING METHODS

Even though the existing methods, CNM [58]¹, CPL [59]², and PPS [20]³, make impressive results and release their codes, the results reported in the original papers are not reproduced properly in our environment. We believe the reason is the different environmental settings, so we adjust hyperparameters of CNM, CPL, and PPS to reproduce the results in our environment. As a result, we are able to successfully reproduce the results reported in the original papers. Tab. 7 shows the results reported in the original papers, the results reproduced in our environmental setting, and the results improved by our proposed method. In the results, our method can boost the performance of existing CNM, CPL, and PPS. This implies that our negative proposals can contribute to high-quality positive proposal generation through contrastive learning.

I EXPERIMENTS FOR LOSS BALANCES

To explore the best loss balances, we conduct experiments with different weights of losses in our method. Tab. 8 provide results of using different weight value λ_1 for the dual-signed cross-entropy loss. The results show that setting $\lambda_1 = 0.03$ yields the best results. High λ_1 causes too high cross-entropy losses of negative proposals in the deconstruction process, resulting in the failure of gradual negative proposal learning. Low λ_1 can not train the negative proposals enough to capture confusing locations. Tab. 9 provide results of different weight value λ_2 for the multi-triplet loss. The results show that setting $\lambda_2 = 1$ yields the best results. High λ_2 causes too much distinction between positive and negative proposals and then the positive proposals only avoid the negative proposals rather than finding query-relevant temporal location. Low λ_2 may deteriorate

¹<https://github.com/minghangz/cnm>.

²<https://github.com/minghangz/cpl>.

³<https://github.com/sunoh-kim/pps>.

Table 7: Performance of the existing methods in the original paper, the existing methods reproduced in our environment, and the existing methods trained with our method. Bold numbers denote the best results.

Method	Charades-STA						ActivityNet Captions					
	R@1		R@5				R@1		R@5			
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
CNM [58]	60.39	35.43	15.45	-	-	-	78.13	55.68	33.33	-	-	-
CPL [59]	66.40	49.24	22.39	96.99	84.71	52.37	82.55	55.73	31.37	87.24	63.05	43.13
PPS [20]	69.06	51.49	26.16	<u>99.18</u>	86.23	53.01	81.84	<u>59.29</u>	31.25	95.28	85.54	71.32
CNM*	61.49	34.86	15.04	-	-	-	78.63	56.36	32.69	-	-	-
CPL*	65.80	49.87	22.36	96.90	83.34	51.14	80.31	52.09	29.68	87.29	62.24	41.69
PPS*	69.00	51.49	<u>26.22</u>	99.15	86.04	53.20	81.78	59.17	31.33	95.19	85.56	71.30
CNM+ours*	64.98	41.61	19.79	-	-	-	80.65	59.07	35.79	-	-	-
CPL+ours*	<u>69.25</u>	<u>52.47</u>	26.09	98.58	<u>89.30</u>	<u>53.55</u>	<u>82.64</u>	59.12	<u>33.56</u>	<u>95.41</u>	<u>85.81</u>	<u>71.39</u>
PPS+ours*	70.74	53.04	26.69	99.24	90.03	53.86	83.56	59.71	33.48	95.50	86.02	71.63

*: the results in our environments.

Table 8: Comparisons of different λ_1 to balance the dual-signed cross-entropy loss on ActivityNet Captions.

Method	λ_1										
	0	0.001	0.01	0.03	0.05	0.1	0.3	0.5	1	3	5
R@1,mIoU	27.15	31.67	33.11	38.49	36.38	36.10	35.45	34.56	33.61	33.19	30.16
R@5,mIoU	49.63	55.73	56.54	61.72	58.48	57.62	56.98	56.14	55.69	55.28	52.27

Table 9: Comparisons of different λ_2 to balance the multi-triplet loss on ActivityNet Captions.

Method	λ_2										
	0	0.001	0.01	0.03	0.05	0.1	0.3	0.5	1	3	5
R@1,mIoU	25.45	26.52	31.19	29.04	32.33	34.40	33.36	35.53	38.49	31.79	28.75
R@5,mIoU	43.71	51.61	54.50	54.61	55.13	55.72	56.39	59.10	61.72	55.16	42.60

the ability of positive proposals to distinguish themselves from negative proposals.

J QUALITATIVE COMPARISONS WITH EXISTING METHODS

For qualitative comparisons, we visualize the predicted temporal boundaries from the given query and video, as shown in Fig. 10. Here, we visualize temporal boundaries of ground truth, our positive proposal, our negative proposals, and positive proposals of other previous methods (*i.e.*, CPL [59] and PPS [20]). We observe that our positive proposals find query-relevant temporal locations closer to the ground truth than other existing methods. Also, our negative proposals can capture positive proposal that fails to find a ground truth temporal boundary in upper-left and upper-right examples of Fig. 10. Our negative proposals can capture a poorly-generated positive proposal while the rule-based negative proposals capture none, which is quantitatively shown in Fig. 6. By capturing various confusing locations including poorly-generated positive proposals, our learning-based negative proposals have higher quality than the rule-based ones, which leads to significant performance improvement as shown in Tab. 4.

Also, we visualize our negative proposals as the training epoch progresses in Fig. 11. Here, we visualize the ground truth temporal boundary, the predicted temporal boundary from positive proposals and negative proposals of our method as the training epoch progresses. The blue texts describe the events that are not relative to the given sentence query, which can be regarded as events for the negative proposals. At the early training stage, our negative proposals can capture events for easy negative, such as “Drawing a game of hopscotch on the ground”. As the training epoch progresses, our negative proposals can capture events for harder negative, such as “The older child stands behind the number one and waves to the camera” and “The lady grabs the rock”. At the late training stage, our negative proposals can capture events for much harder negative, such as “The older child hops all the way through the game”. By capturing confusing locations described by the various events, our negative proposals can achieve higher performance than the rule-based ones in Tab. 4.

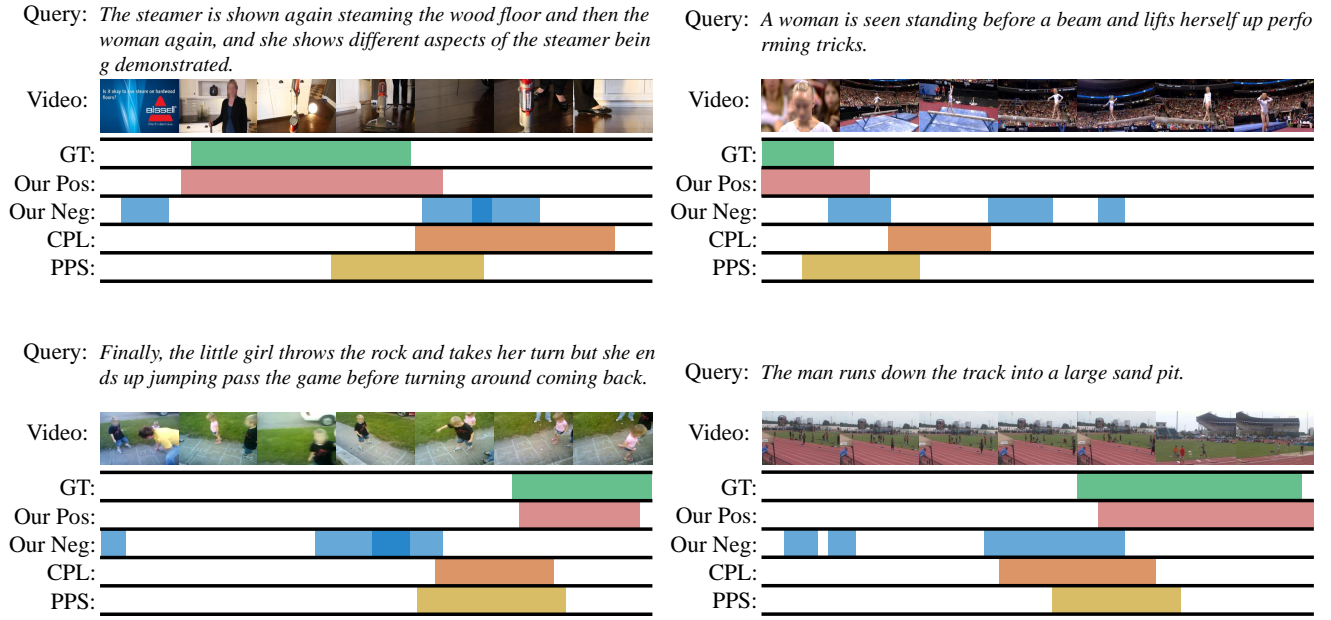


Figure 10: Qualitative results of predicted temporal boundaries. We visualize the ground truth temporal boundary (Green), the predicted temporal boundary from the positive proposal of our method (Red), the multiple temporal boundaries from multiple negative proposals of our method (Blue), the predicted temporal boundary from the positive proposal of other methods, i.e., CPL (Orange) and PPS (Yellow).

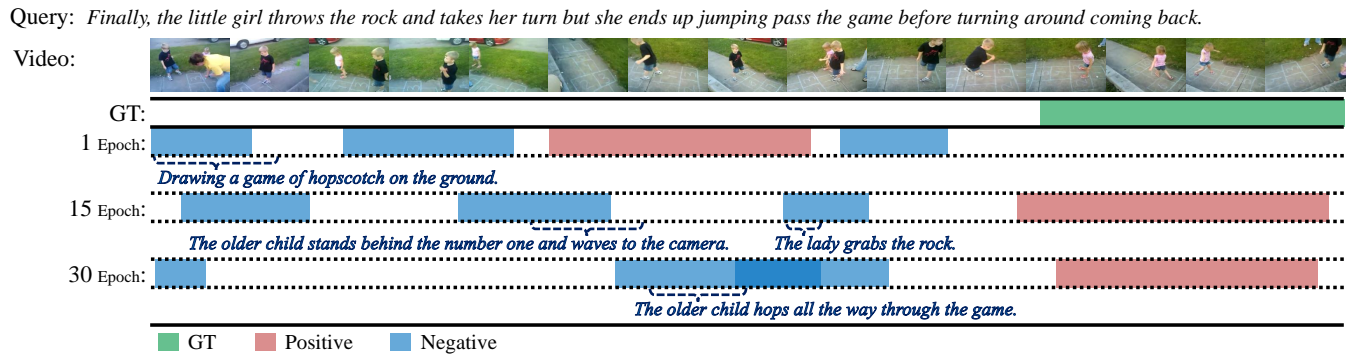


Figure 11: Qualitative results of our negative proposals changing from easy to hard ones. We visualize the ground truth temporal boundary (Green), positive proposals (Red), and negative proposals (Blue) as the training epoch progresses. The blue texts describe the events that are not relative to the given sentence query, which can be regarded as events for the negative proposals.