

# 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

## A. Visualizations

In this section, we present visualizations of some false negative moments that have been eliminated by our Semantic Boundary Detection (SBD). Figure 1 illustrates the results, where the first column displays the corresponding video of the original single target moment, the second column shows the false negative moment found within the same video, and the third column shows the false negative moment discovered in other videos of the same batch. It is evident that the identified false negative moments align with the query description. Figure 2 illustrates the prediction results for some additional multi-target samples. Our method shows improved precision in finding other false negative moments compared to our baseline [15], due to the utilization of multi-target training enabled by Semantic Fusion Augmentation (SFA) and the improved embedding space achieved through Intra-Video Contrastive Loss and SBD.

## B. Implementation Details

The hyperparameters for each dataset and their corresponding video encoders are shown in Table 1. Other implementation details will be provided in this section.

For Charades and ActivityNet, we adopt the approach proposed in previous work [15] to generate the 2D proposal map from the video feature  $v$ . Similarly, for QVHighlights, we use the same configuration as ActivityNet to generate the 2D proposal map. However, instead of using ConvNet [15] to aggregate the features of the proposal, we use ResNet-18 [4] as the proposal encoder for all datasets. Our ResNet-18 has some modifications: All convolutions in ResNet-18 are replaced by masked convolutions [23] used in ConvNet. The first convolution has a kernel size of 5, the max-pooling layer is removed, the stride sizes of all convolutions are set to 1, and the channel sizes of all convolutions are set to 256 for Charades and ActivityNet, and 512 for QVHighlights. As for the language encoder, we employ DistilBert [12] provided by HuggingFace [16]. We use the pre-trained model "distilbert-base-uncased" following [15]. Following the previous work [15, 23], our

predictions are generated using non-maximum suppression, which removes overlapped proposals with low confidence.

To prevent GPU memory leakage, we limit the number of queries per video to a maximum of 7, resulting in a total of at most  $B \times 7$  video query pairs. Before applying SFA, the candidate target moment has a probability of  $p_d$  being down-sampled to half the length by choosing the odd video sequence features along the time axis. After down-sampling, the candidate target moment has the probability of  $p_a$  being mixed with another random moment. If it is not possible to find a non-overlapping moment, the SFA process is skipped. The mixup ratio between the target moment and the random moment is set to 9 : 1 for all experiments to ensure that the augmented moment still shares most of the semantics with the original target moment.

We utilize AdamW [10] as an optimizer. In all experiments except QVHighlights, the contrastive loss weights  $\lambda_{\text{inter}}$  and  $\lambda_{\text{intra}}$  are decayed by a factor of 0.01 at the beginning of the 7th epoch. The threshold for non-maximum suppression is set to 0.5 in all experiments. The experiments involving ActivityNet with the I3D feature are carried out using 2×NVIDIA RTX 3090. For all other experiments, a single NVIDIA RTX 3090 is utilized.

For Charades and ActivityNet, we report the best performance with respect to  $R@1, \text{IoU}=0.7 + R@(5,5), \text{IoU}=0.5$ . In the case of QVHighlights, we identify the best model based on the assessed mAP@avg on the validation set. In addition, we employ this model to generate predictions for submission to the QVHighlights evaluation server.

## C. Ablation Details

For Table 8 in the paper, we evaluated various settings. This involved testing fixed thresholds of  $\{0.5, 0.7, 0.9\}$ , as well as two types of predefined acceptance rate schedulers: linearly increasing from 0 to 1 throughout training, and stepping from 0 to 1 at the midpoint of the training epoch. The reported results are based on the optimal results in terms of  $R@1, \text{IoU}=0.7 + R@(5,5), \text{IoU}=0.5$  for each combination.

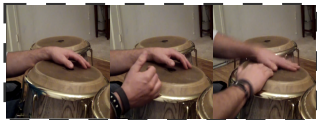
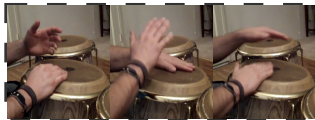




Query	Positive Sample	False Negative (Same Video)	False Negative (Other Video)
A man is explaining how to play the drums on bongos.	 Video ID: v_qsTCTQo-wl8 10s - 36s	 Video ID: v_qsTCTQo-wl8 137s - 161s	 Video ID: v_u35hesPTsNE 15s - 52s
Person eating some food.	 Video ID: DWBS3 0s - 4s	 Video ID: DWBS3 12s - 21s	 Video ID: WZVHJ 18s - 25s

Figure 1. False negative samples from the ActivityNet training set (first row) and the Charades training set (second row).

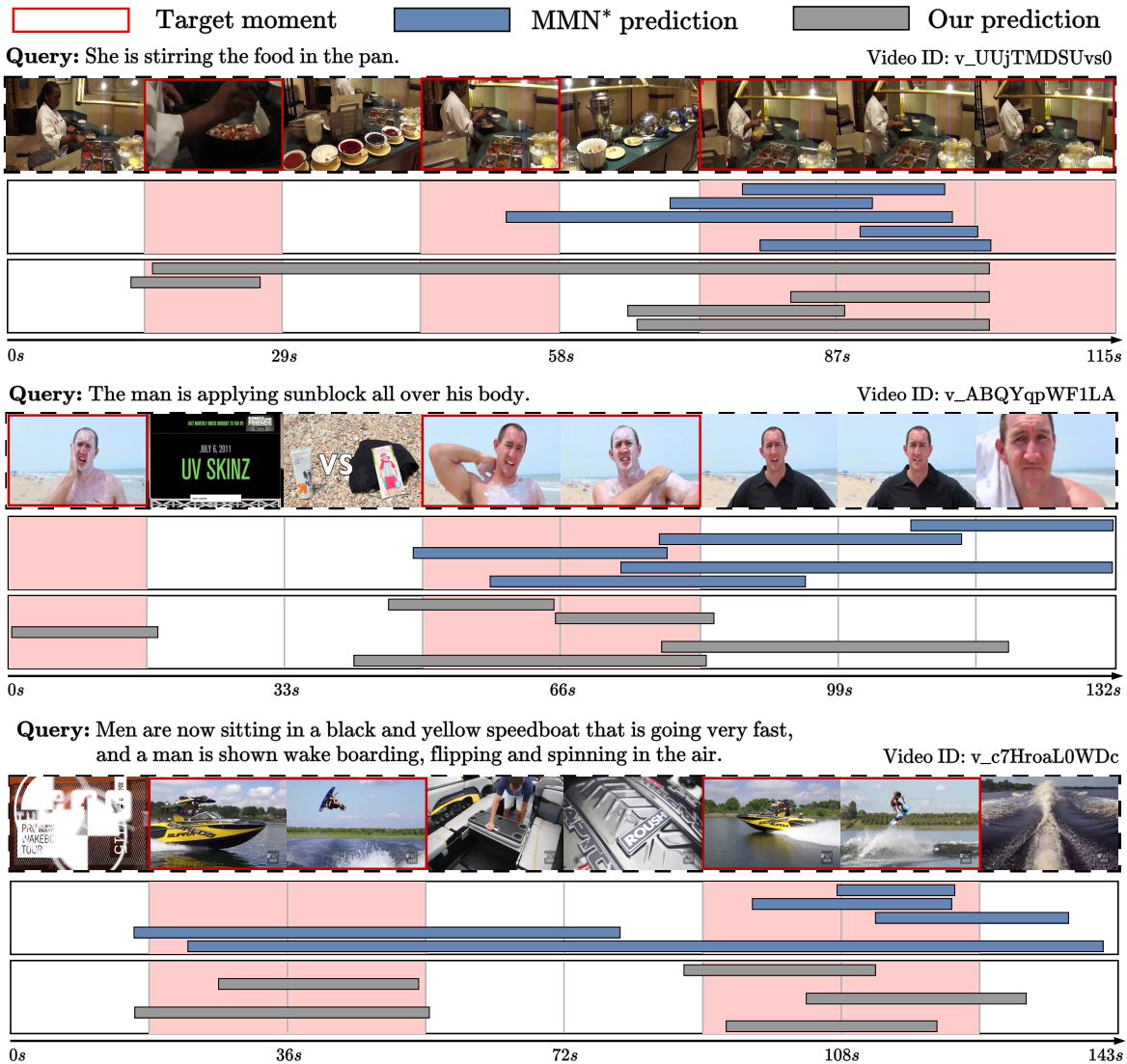


Figure 2. Visualization of some prediction results. It should be noted that MMN\* is implemented in our code base.

Dataset	Charades-STA			ActivityNet		QVHighlights
Feature	VGG	C3D	I3D	C3D	I3D	SlowFast + CLIP
Model lr	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
DistilBert lr	1e-5	1e-5	1e-5	5e-5	5e-5	1e-5
$\lambda_{\text{inter}}$	0.05	0.1	0.1	0.05	0.05	0.5
$m_{qv}$	0.3	0.3	0.3	0.3	0.3	0.3
$\tau_{qv}$	0.1	0.1	0.1	0.1	0.1	0.1
$m_{vq}$	0.3	0.3	0.3	0.3	0.3	0.3
$\tau_{vq}$	0.1	0.1	0.1	0.1	0.1	0.1
$\lambda_{\text{intra}}$	0.01	0.05	0.1	0.01	0.01	0.5
$m_{vv}$	0.2	0.2	0.2	0.2	0.2	0.0
$\tau_{vv}$	0.05	0.05	0.05	0.05	0.05	0.05
$N$	16	16	16	64	64	64
Batch size	48	48	48	24	16	24
$\alpha$	1.0	0.25	1.0	1.0	0.5	1.0
Bert fire start [15]	1	4	1	5	5	1
Milestone [15]	9	8	8	8	8	$\times$
$d_e$	256	256	256	256	256	512
$p_a$	0.25	0.25	0.25	0.25	0.25	0.25
$p_d$	0.5	0.5	0.0	1.0	0.0	0.0
Dual space [15]	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$

Table 1. Hyper-parameters for each dataset and the corresponding video feature.

Level	Augmentation	R@1	R@5	R@(5,5)
		IoU=0.7	IoU=0.7	IoU=0.7
Frame	Cutmix	25.82	57.37	28.90
Frame	Mixup	<b>27.59</b>	<u>57.42</u>	28.68
Feature	Cutmix	26.94	57.01	<u>29.10</u>
Feature	Mixup	<u>27.05</u>	<b>58.35</b>	<b>29.72</b>

Table 2. Ablation of augmentation on Charades using VGG features. Please note that the pre-trained VGG weights for Charades are unavailable. Therefore, we utilize the built-in PyTorch weights for the experiments in this table.

For Table 2, we examined various configurations of mixup and cutmix operations. Regarding mixup, we assessed mixing rates such as  $\{1 : 9, 3 : 7, 5 : 5\}$  between background features (frames) and target features (frames). In the case of cutmix at the feature level, we directly replaced  $\alpha\%$  of the elements in the background features with the corresponding target features. For cutmix at the frame level, we directly substituted a random square area of size  $\alpha^2$  in the original frame with the corresponding area from the target frame. We experimented with different values of  $\alpha$ , specifically in the range  $\{0.5, 0.7, 0.9\}$  and reported the best result in terms of  $R@1, \text{IoU}=0.7 + R@(5,5), \text{IoU}=0.5$  for each combination.

## D. Illustration of Challenge

Figure 3 illustrates the potential challenge of achieving a simultaneous improvement on both multi-target metric

$R@(5,5)$  and single-target metric  $R@1$ .

## E. Supplemental Experiments

Table 4, 5 present the ablation results of Charades with I3D feature and ActivityNet with I3D feature.

We present the evaluation results of our SFABD on commonly used datasets with other video encoders in Tabs. 3, 6 and 7 to facilitate future comparisons. It can be seen that our SFABD also achieves comparable results in Charades and ActivityNet using other video encoders.

Table 8 shows the result of using different proposal encoders in ActivityNet with the C3D feature. Similarly to the results of Table 7 in our paper, the performance improvement of our SFABD compared to MMN is greater when ResNet-18 is used. This indicates that our method has the potential to yield even greater performance boosts utilizing a better proposal encoder.

Due to the submission limits of the QVHighlights evaluation server, we conducted our ablation study for QVHighlights only on the validation set. As shown in Table 9, our SFABD outperforms previous methods by a significant margin in mAP@avg. This aligns with the results obtained on the testing set (Table 3 of our paper). Unlike [11], we do not utilize audio as an additional input source. We leverage the multi-target information through intra-video contrastive learning, which is the key difference between our method and previous self-attention-based approaches. Such a self-attention-based approach lacks intra-video supervision dur-

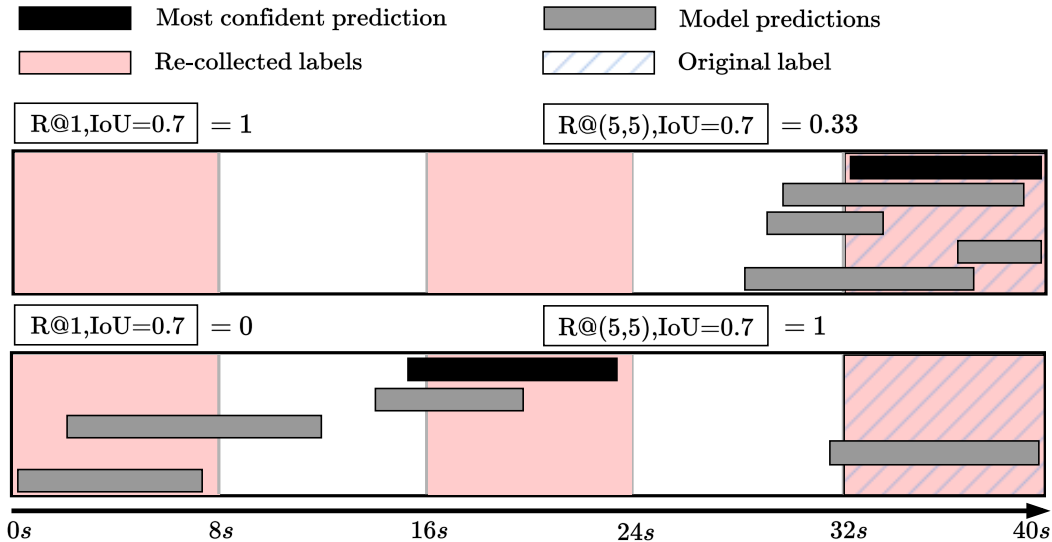


Figure 3. This figure shows an imaginary sample with its original label, re-collected labels and two contrasting imaginary prediction results of a VMR model similar to the situation of the first example of Figure 2. The first row illustrates the imaginary prediction results of a model that learns a single-target prediction bias due to single-target training. Although  $R@1$  is high,  $R@(5,5)$  remains low, highlighting that only evaluating the single-target performance with the original label neglects the multi-target performance required for multi-target VMR. The second row shows the imaginary prediction results of the model that has undergone multi-target training, a more diverse prediction results can be achieved. Although  $R@(5,5)$  is high,  $R@1$  remains low because the most confident prediction does not align with the original label. In order to achieve simultaneous improvement in both  $R@(5,5)$  and  $R@1$ , a model not only needs to generate diverse predictions that capture all target moments, but also needs to make sure that the most confident prediction aligns with the original label. However, since the multi-target labels should hold equal priority during multi-target training, demanding such alignment property is unreasonable.

Method	C3D video features				I3D video features			
	R@1		R@5		R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
DRN [19]	45.40	26.40	<u>88.01</u>	55.38	53.09	31.75	89.06	60.05
VSLNet [20]	<u>47.31</u>	<u>30.19</u>	-	-	-	-	-	-
MS-2D-TAN [22]	41.10	23.25	81.53	48.55	56.64	36.21	89.06	61.13
DTG [25]	-	-	-	-	<u>60.19</u>	39.38	87.53	66.91
DTG-SPL [24]	-	-	-	-	60.05	40.13	87.34	67.12
BMRN [13]	45.93	28.37	<b>89.12</b>	<u>57.19</u>	<b>63.09</b>	<b>42.46</b>	<b>92.62</b>	<u>67.65</u>
SFABD (Ours)	<b>47.86</b>	<b>30.51</b>	85.48	<b>59.96</b>	60.09	<u>40.21</u>	<u>90.36</u>	<b>68.65</b>

Table 3. Evaluation results on Charades with the C3D feature and the I3D feature.

ing the video sequence encoding process, which results in a degradation of the model’s capability. Furthermore, Table 10 shows the method ablation in the QVHighlights validation set. We observe that even though QVHighlights already contains many multi-target samples, using augmentation can still lead to a slight improvement in  $mAP@avg$ . We attribute the slight improvement to the augmentation of the minority single-target samples in QVHighlights. By augmenting these single-target samples, we can further increase the diversity of multi-target samples. Additionally,

we find that incorporating intra-video contrastive loss leads to significant performance gain. This result substantiates our early statement that utilizing the information in multi-target labels is crucial to achieve better multi-target VMR performance. We also find that SBD does not contribute much in QVHighlights, and we believe the main reason is that QVHighlights is a dataset with high-quality multi-target labels and diverse queries. Consequently, the probability of encountering false negative moments in a batch is significantly lower.

$\mathcal{L}_{inter}$	SFA	$\mathcal{L}_{intra}$	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
✓				58.57	38.90	88.56	68.29	63.26	35.18
✓	✓			58.13	38.33	88.59	66.58	63.90	36.12
✓			✓	58.81	39.88	89.40	68.65	62.92	35.70
✓	✓	✓		<u>59.25</u>	<u>39.53</u>	<u>89.49</u>	<b>69.30</b>	<u>63.40</u>	<b>37.10</b>
✓	✓	✓	✓	<b>60.09</b>	<b>40.21</b>	<b>90.36</b>	<u>68.65</u>	<b>65.14</b>	<u>36.16</u>

Table 4. Method Ablation on Charades-STA with the I3D Feature.

$\mathcal{L}_{inter}$	SFA	$\mathcal{L}_{intra}$	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
✓				48.27	29.92	80.23	66.58	59.55	42.41
✓	✓			48.90	30.75	79.49	63.90	<u>60.65</u>	42.66
✓			✓	48.46	30.21	81.01	<b>67.12</b>	<u>59.77</u>	<u>42.90</u>
✓	✓	✓		<u>49.13</u>	<u>30.84</u>	<u>81.02</u>	65.70	60.37	<u>42.76</u>
✓	✓	✓	✓	<b>49.22</b>	<b>30.97</b>	<b>81.03</b>	<u>66.81</u>	<b>60.79</b>	<b>43.77</b>

Table 5. Method ablation on ActivityNet with the I3D feature.

Method	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
MCN [1]	17.46	8.01	48.22	26.73
DRN [19]	42.90	23.68	<u>87.80</u>	54.87
2D-TAN [23]	39.70	23.31	80.32	51.26
MS-2D-TAN [22]	45.65	27.20	86.72	56.42
CBLN [8]	43.67	24.44	<b>88.39</b>	56.49
FVMR [3]	42.36	24.14	83.97	50.15
MMN [15]	47.45	27.15	83.82	<u>58.09</u>
QD-DETR [11]	<b>52.77</b>	<u>31.13</u>	-	-
SFABD (Ours)	<u>50.23</u>	<b>31.38</b>	85.62	<b>61.07</b>

Table 6. Evaluation results on Charades with the VGG feature.

Method	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
MCN [1]	21.36	6.43	53.23	29.70
CTRL [2]	29.01	10.34	59.17	37.54
QSPN [17]	33.26	13.43	62.39	40.78
SCDM [18]	36.75	19.86	64.99	41.53
DRN [19]	45.45	24.36	77.97	50.30
MSA [21]	48.02	<b>31.78</b>	78.02	63.18
2D-TAN [23]	44.51	26.54	77.13	61.96
CPNet [7]	40.56	21.63	-	-
CBLN [8]	48.12	27.60	79.32	63.41
FVMR [3]	45.00	26.85	77.42	61.04
MMN [15]	48.04	29.68	79.49	<u>65.12</u>
TACI [14]	45.50	27.23	-	-
MS-2D-TAN [22]	46.16	29.21	78.80	60.85
STCM-Net [5]	46.23	29.04	78.43	63.46
BMRN [13]	<u>48.47</u>	<u>31.15</u>	<b>81.37</b>	64.44
SFABD (Ours)	<b>49.00</b>	31.10	<u>80.82</u>	<b>65.55</b>

Table 7. Evaluation results on ActivityNet with the C3D feature.

Method	Proposal Encoder	R@1		R@5	
		IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
MMN	ConvNet	48.59	29.26	79.50	64.76
MMN <sup>†</sup>	ConvNet	47.65	29.20	79.62	<b>65.95</b>
SFABD	ConvNet	48.80	30.62	79.94	64.95
MMN <sup>†</sup>	ResNet-18	<u>48.69</u>	<u>29.39</u>	<u>80.20</u>	65.06
SFABD	ResNet-18	<b>49.00</b>	<b>31.10</b>	<b>80.82</b>	<u>65.55</u>

Table 8. Evaluation results on ActivityNet with the C3D feature and different proposal encoders. ConvNet is the encoder officially used by MMN [15]. Note that <sup>†</sup> denotes that the implementation is based on our code base.

Method	mAP@0.5	mAP@0.75	mAP@avg
momentDETR [6]	-	-	36.30
UMT <sup>†</sup> [9]	-	-	38.59
QD-DETR <sup>†</sup> [11]	<u>62.23</u>	<u>41.82</u>	<u>41.22</u>
SFABD (Ours)	<b>63.09</b>	<b>45.99</b>	<b>45.65</b>

Table 9. Evaluation results on the QVHighlights validation set. The symbol <sup>†</sup> indicates that the source feature includes video and audio. Otherwise, the input feature consists of video only.

$\mathcal{L}_{inter}$	SFA	$\mathcal{L}_{intra}$	SBD	mAP@avg
✓				43.79
✓	✓			44.17
✓		✓		44.85
✓			✓	44.40
✓	✓	✓		<u>45.64</u>
✓	✓	✓	✓	<b>45.65</b>

Table 10. Method ablation on QVHighlights validation set.



## 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. 5
- [2] 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. 5
- [3] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021. 5
- [4] 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. 1
- [5] Zixi Jia, Minglin Dong, Jingyu Ru, Lele Xue, Sikai Yang, and Chunbo Li. Stem-net: A symmetrical one-stage network for temporal language localization in videos. *Neurocomputing*, 471:194–207, 2022. 5
- [6] 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. 5
- [7] 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. 5
- [8] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021. 5
- [9] 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
- [10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1
- [11] 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. 3, 5
- [12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*, 2019. 1
- [13] 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. 4, 5
- [14] Jungkyoo Shin and Jinyoung Moon. Learning to combine the modalities of language and video for temporal moment localization. *Computer Vision and Image Understanding*, 217:103375, 2022. 5
- [15] 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. 1, 3, 5
- [16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 1
- [17] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019. 5
- [18] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019. 5
- [19] 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. 4, 5

- [20] 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. 4
- [21] 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. 5
- [22] 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. 4, 5
- [23] 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. 1, 5
- [24] Hao Zhou, Chongyang Zhang, Yanjun Chen, and Chuanping Hu. Towards diverse temporal grounding under single positive labels. *arXiv preprint arXiv:2303.06545*, 2023. 4
- [25] 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. 4