Appendix for "*Query-Dependent Video Representation for Moment Retrieval and Highlight Detection*"

WonJun Moon[1,*], Sangeek Hyun[1,*], SangUk Park[2], Dongchan Park[2], Jae-Pil Heo[1,⋆]
[1]Sungkyunkwan University, [2]Pyler
{wjun0830, hsi1032, jaepilheo}@g.skku.edu, {psycoder, cto}@pyler.tech

1. Training Details

In this section, we elaborate on the implementation details and hyperparameters used for experiments in the main manuscript. To unify configurations across all experiments, our encoder composes of 4 layers of transformer block (2 cross-attention layers and 2 self-attention layers) whereas there are only 2 layers in the decoder (For HD dataset, i.e., TVSum, we only use encoding layers). We set the hidden dimension of transformers as 256, and use the Adam optimizer with a weight decay of 1e-4. Besides, we set the temperature of a scaling parameter τ for contrastive loss as 0.5 for all experiments. Loss balancing parameters are $\lambda_{\text{margin}} = 1$, $\lambda_{\text{cont}} = 1$, $\lambda_{L1} = 10$, $\lambda_{\text{gIoU}} = 1$, $\lambda_{\text{CE}} = 4$ and $\lambda_{\text{neg}} = 1$, unless otherwise mentioned. Additionally, we use the PANN [5] model trained on AudioSet [3] to extract audio features[1] for experiments with the audio modality.

Other configurations are described as follows:

QVHighlight. We use video features extracted from both pretrained SlowFast [2] (SF) and CLIP encoder [8], and text embeddings from CLIP, following the Moment-DETR. We train QD-DETR for 200 epochs with a batch size of 32 and a learning rate of 1e-4.

Charades-STA. We utilize official VGG [9] features with GloVe [7] text embedding. To compare with additional baselines, we also test our model on pretrained C3D [10], SlowFast and CLIP for video features with CLIP text embedding. Specifically, we utilize pre-extracted features provided by other baselines repositories: UMT[1], VSLNet[2] and Moment-DETR[3]. We train ours for 100 epochs with a batch size of 8 and a learning rate of 1e-4.

TVSum. I3D [1] features pretrained on Kinetics-400 [4] are utilized as a visual one, and CLIP features are used for the text embedding. Following the most recent work [6], we train our model for 2000 epochs with a learning rate of 1e-3. The batch size is set to 4.

[1]https://github.com/TencentARC/UMT
[2]https://github.com/IsaacChanghau/VSLNet
[3]https://github.com/jayleicn/moment_detr

Table 1. Experimental results on QVHighlights.

		MR					HD	
		R1		mAP			>= Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Performances with respect to query length; S: # words $\leq$ 8, M: 8 $<$ # words $\leq$ 13, L: 13 $<$ # words								
S	M-DETR	51.82	34.49	51.48	29.48	29.43	37.11	59.27
	QD-DETR	63.95	48.18	61.18	40.93	40.23	38.67	63.60
M	M-DETR	57.47	39.22	57.41	33.43	34.73	37.49	56.26
	QD-DETR	65.91	51.43	65.48	45.54	44.46	40.07	62.90
L	M-DETR	49.35	32.90	52.89	29.14	30.54	35.95	55.16
	QD-DETR	57.42	40.32	61.03	37.67	38.56	39.24	61.29

2. Further study on model performance on varying lengths of the query.

As discussed in the limitation, the performance of QD-DETR may depend on the quality of provided ground truth text descriptions. Yet, this does not imply the QD-DETR's vulnerability against commonly used meaningless words in text descriptions. As we think the queries with longer lengths may have a higher chance of including noisy texts, we divide the validation set into 3 groups each with long-, medium-, and short-length queries, and report the query-length-wise performances of QD-DETR in Tab. 1. As shown, QD-DETR works well regardless of the query length, showing [36.7, 28.0, 26.3%] and [7.3, 11.8, 11.1%] improvements in mAP each for MR and HD with [Short, Medium, Long] queries. This study implies that while irrelevant (wrong) text descriptions for video contexts can degrade the effectiveness of QD-DETR, QD-DETR is robust against meaningless words that are commonly present in text queries.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[3] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 1

[6] 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

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[8] 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. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1