APPENDIX

A ADDITIONAL INFORMATION ON TRAINING

A.1 The 1-st Stage Search Engine

In order to make CONQUER independent of HERO (in case of experiment bias), we indeed use a simplified version of CONQUER as the 1-st stage search engine. Referring to Fig. 2(a), we still use the MMT and transformer to compute $\widehat{\mathbf{V}}$, $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{Q}}$ for QDF. NetVLAD is used to derive the fusion weight μ^v, μ^t. Same as XML [19], we also learn to aggregate $\widehat{\mathbf{Q}}$ into two vectors as query vectors for visual and textual modalities, respectively. The vectors are directly used to compute the video similarities with $\widehat{\mathbf{V}}$, $\widehat{\mathbf{S}}$. The two similarities scores are further weighted sum using the fusion weights μ^v, μ^t. Note that QAL and ML heads are not used in this simplified CONQUER.

A.2 Negative Video Sampling

During training, we sample videos of a mini training batch up to a search depth of $d = p + x$, where p is the rank of the ground-truth video, and x is the extension range. Fig. 4 shows the VCMR result sum curves for different similarity scoring functions when the negative video number is set as 3. As we can see, $x = 500$ achieves the optimal performance and is the final adopted setting in this work. Compared with the performance of $x = 500$, the performances of randomly sampled negative videos (see $x = 17,435$) are always inferior. It indicates the importance to sample top-ranked videos as negative videos.

B THE MOMENT LENGTH CONSTRAINS

Same as XML [19] and HERO [20], we assume the knowledge of moment length for pruning short and lengthy moments. We use $L_{min} = 1$ and $L_{max} = 24$, instead of $L_{min} = 2$ and $L_{max} = 16$ as suggested by HERO for TVR dataset. Table 8 provides the experimental results if we set the moment length constrains as HERO, i.e., $L_{min} = 2$ and $L_{max} = 16$. As shown in Table 8, CONQUER still outperform HERO and XML significantly by using this setting. Note that the setting ($L_{min} = 3$ and $L_{max} = 7$) on DiDeMo is the same for both HERO and CONQUER.

C MORE RESULTS ON TVR

Table 9 provides the full set of results comparing CONQUER with XML, HAMMER and HERO. The results include performance for VCMR, VR and SVMR.

Table 8: Results of CONQUER if using the setting of $L_{min} = 2$ and $L_{max} = 16$ suggested by HERO for result pruning.

Model	VCMR				SVMR IoU=0.7	
	R1	R5	R10	R100	R1	R5
XML [19]	2.62	6.39	9.05	22.47	13.89	31.11
HERO [20]	5.13	12.24	16.26	24.56	15.36	32.27
CONQUER (general)	**7.25**	16.49	**21.66**	**33.88**	**21.87**	46.18
CONQUER (disjoint)	6.89	**16.88**	22.38	33.29	21.03	**46.98**
CONQUER (exclusive)	6.79	16.35	21.57	33.8	19.55	43.83

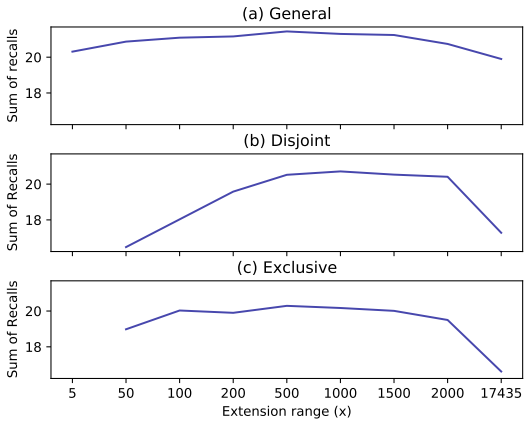

Figure 4: Effects of negative video sampling. The y-axis shows the sum of R@1 for both IoU={0.5, 0.7} and x-axis is the extension range of search depth.

Table 9: Full results on TVR validation set.

Model	VCMR IoU=0.7				VR		SVMR IoU=0.7	
	R1	R5	R10	R100	R1	R5	R1	R5
XML [19]	2.62	6.39	9.05	22.47	16.08	37.92	13.89	31.11
HAMMER	5.13	-	11.38	16.71	-	-	-	-
HERO [20]	5.13	12.24	16.26	24.56	29.01	52.82	15.36	32.27
CONQUER (general)	**7.76**	17.22	22.49	**35.17**	29.01	52.82	**22.84**	47.72
CONQUER (disjoint)	7.18	**17.4**	**23.0**	33.94	26.67	55.89	21.84	**48.63**
CONQUER (exclusive)	7.02	17.03	22.44	34.69	**29.29**	**56.15**	20.1	45.12