Appendix

A. Effectiveness of SA-GCL and DGA

To further analyze the effectiveness of SA-GCL and DGA, we provide more detailed experimental results on ActivityNet Captions and TACoS datasets as shown in Table 1 and Table 2. Following the main manuscript, we regard the simplified implementation of SA-GCL as a baseline. After being equipped with the complete SA-GCL, our model achieves significant improvements on both ActivityNet Captions and TACoS. This phenomenon demonstrates that sampling enough positive moments for contrastive learning is of great importance. Additionally, we further incorporate the DGA module for alleviating the annotation bias and modeling complex target moments. Since the ActivityNet Captions dataset has a large number of complex query sentences consisting of multiple events, D3G obtains notable performance gains on ActivityNet Captions(*e.g.* 9.03% at R@5 IoU=0.7). However, TACoS is still challenging for D3G due to the dense distributions of target moments.

Module		R@1		R@5	
SA-GCL	DGA	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
✓†		0.83	0.28	1.78	0.58
✓		32.65	16.00	65.48	43.44
✓	✓	36.68	18.54	74.21	52.47

Table 1. Effectiveness of SA-GCL and DAG in D3G on ActivityNet Captions. ✓† denotes an simplified implementation of SA-GCL.

Module		R@1		R@5	
SA-GCL	DGA	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
✓†		2.97	0.37	5.40	1.10
✓		11.95	4.20	29.07	10.30
✓	✓	12.67	4.70	31.34	12.35

Table 2. Effectiveness of SA-GCL and DAG in D3G on TACoS. ✓† denotes an simplified implementation of SA-GCL.

B. Effect of different hyper-parameters

In this section, we investigate the effect of two critical hyperparameters on ActivityNet Captions and TACoS datasets. As shown in Figure 1 and Figure 2, we report the changes in performance at four metrics. As for top-k, the performance increases dramatically as the k increases. However, the performance gradually achieves saturation after the k reaches 15. We finally select $k = 20$ for both ActivityNet Captions and TACoS. As for σ, the ActivityNet

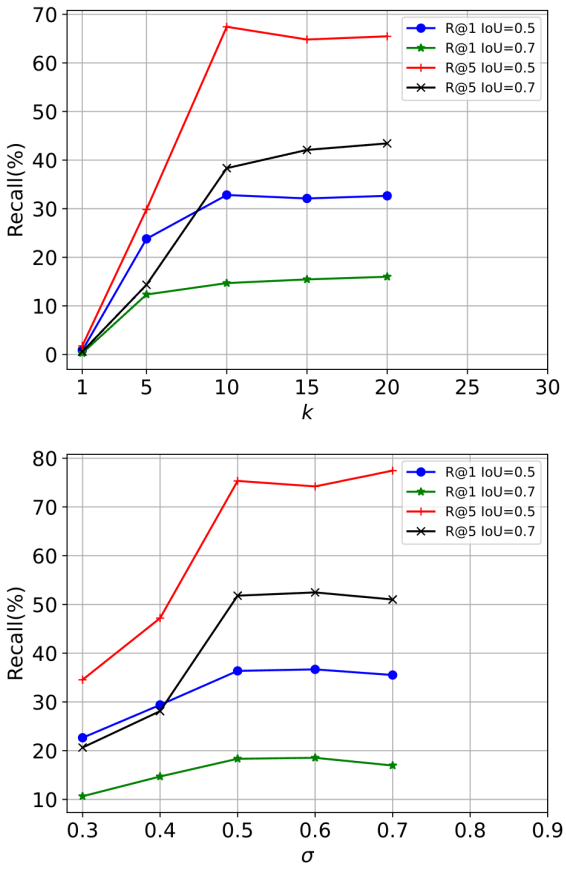

Figure 1. Effect of top-k and σ on ActivityNet Captions dataset.

Captions dataset tends to prefer large values while small values are more suitable for the TACoS dataset. This is because the former contains a large number of long target moments while the latter contains numerous short target moments. As shown in Figure 1 and Figure 2, we eventually select $\sigma = 0.6$ and $\sigma = 0.2$ for ActivityNet Captions and TACoS for optimal performance, respectively.

C. Qualitative Analysis

In this section, we provide more qualitative examples from the test split of the Charades-STA dataset, ActivityNet Captions dataset, and TACoS dataset. For each video, we select two queries for analysis. As shown in Figure 3 (a), D3G locates the target moment accurately while ViGA ignores the reason at the front of the target moment, given Query 1. However, D3G is inferior to ViGA in some cases such as Query 2. As for complex queries in ActivityNet Captions, D3G still localizes a moment with a large overlap with the target moment. Since sentence-level features may lose some information about specific events, D3G cannot perceive accurate boundaries for some complex queries, such as Figure 3 (b) Query 2. It is expected to explore event-

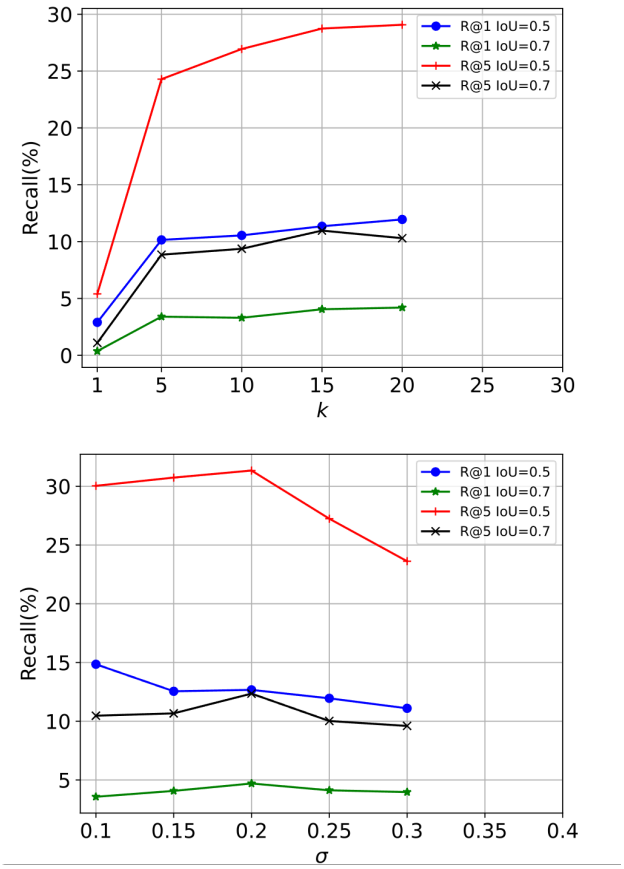

Figure 2. Effect of top-k and σ on TACoS dataset.

level features for queries consisting of multiple events in the future. TACoS is the most challenging dataset, where the videos have long durations and contain a large number of moment-sentence pairs. As shown in Figure 3 (c), we observe that D3G fails to locate a simple query of short duration from the long video, given Query 1. However, D3G accurately locates the target moment of long duration given Query 2. Note that D3G well attends to the number "the last two" of the query while ViGA fails to attend to such information and locates irrelevant moments. As observed in Figure 3, D3G is superior to ViGA, which is consistent with the experimental results in the main manuscript. However, D3G still has some limitations and needs to be improved in the future.

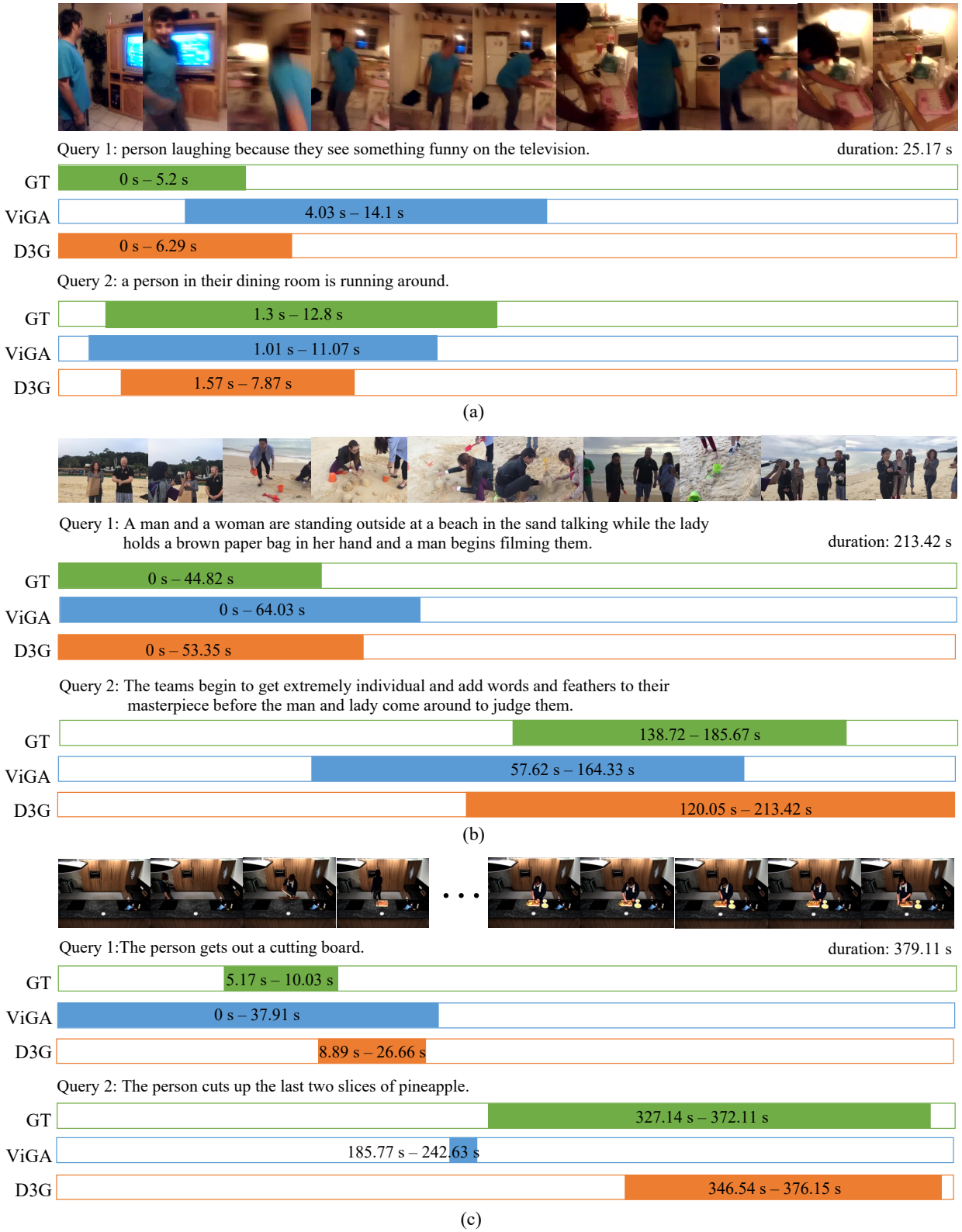

Figure 3. Qualitative examples of top-1 predictions. (a), (b) and (c) is from the Charades-STA dataset, the ActivityNet Captions and the TACoS dataset, respectively. GT indicates the ground truth temporal boundary.