Supplementary Material:
Support-Set Based Cross-Supervision for Video Grounding

Xinpeng Ding[1,2], Nannan Wang[1], Shiwei Zhang[2], De Cheng[1], Xiaomeng Li[3],
Ziyuan Huang [4], Mingqian Tang[2], Xinbo Gao[5]
[1]Xidian University, [2]Alibaba Group, [3]The Hong Kong University of Science and Technology,
[4]National University of Singapore, [5]Chongqing University of Posts and Telecommunications
xpding.xidian@gmail.com, {nnwang,dcheng}@xidian.edu.cn, eexmli@ust.hk
{zhangjin.zsw, mingqian.tmq}@alibaba-inc.com, ziyuan.huang@u.nus.edu, gaoxb@cqupt.edu.cn

In this supplementary material, we present more quantitative results of prediction and ablation studies.

1. Qualitative evaluation

Fig. 1 shows the comparison of the predicted time intervals of the baseline model (2D-TAN) and Ours (2D-TAN + SS). It is clear that plugging our support-set based supervision to the baseline model, the predicted time intervals are more accurate.

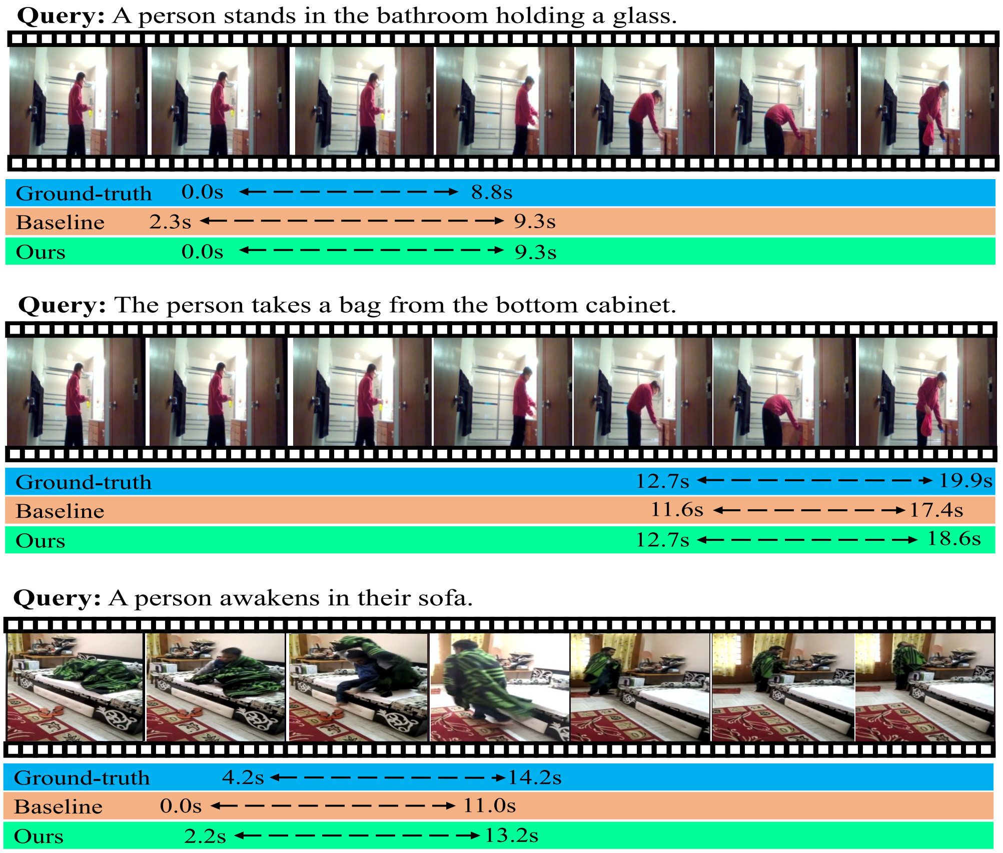
Query: A person stands in the bathroom holding a glass.
Ground-truth 0.0s ← → 8.8s
Baseline 2.3s ← → 9.3s
Ours 0.0s ← → 9.3s
Query: The person takes a bag from the bottom cabinet.
Ground-truth 12.7s ← → 19.9s
Baseline 11.6s ← → 17.4s
Ours 12.7s ← → 18.6s
Query: A person awakens in their sofa.
Ground-truth 4.2s ← → 14.2s
Baseline 0.0s ← → 11.0s
Ours 2.2s ← → 13.2s

Figure 1. Qualitative results of predicted time intervals.

Table 1. Ablation study of different positive and negative sets of the ground-truth clip based supervision on the Charades-STA dataset.

$\mathcal{P}$	$\mathcal{N}$	$Rank1@$		$Rank5@$	
		0.5	0.7	0.5	0.7
GT	Non-GT + V^o	**54.77**	**31.63**	86.28	55.07
GT	Non-GT	53.56	30.55	85.87	54.35
GT	V^o	54.24	30.91	86.16	55.10
V^i	V^o	54.37	31.08	**86.69**	**55.54**

2. Ablation Study

We present different positive and negative sets of the ground-truth clip based supervision (GTC) on the Charades-STA dataset in Table 1. 'GT' indicates ground-truth clips and 'Non-GT' indicates the non-ground-truth clips. 'V^i' are clips in the video corresponding to the text query and 'V^o' are clips in the other videos in the batch. Considering all clips in the video corresponding to the text query as positive sets would have higher mAP value at $Rank5$, while the baseline GTC ($\mathcal{P}$ = GT, $\mathcal{N}$ = Non-GT+V^o) have a higher performance at $Rank1$.