Supplemental Material: *Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention*

Cristian Rodriguez-Opazo[1,2] Edison Marrese-Taylor [3] Fatemeh Sadat Saleh[1,2]
Hongdong Li[1,2] Stephen Gould[1,2]

Australian National University [1], Australian Centre for Robotic Vision (ACRV) [2]
{cristian.rodriguez, fatemehsadat.saleh, hongdong.li, stephen.gould}@anu.edu.au
Graduate School of Engineering, The University of Tokyo [3]
emarrese@weblab.t-utokyo.ac.jp

1. Charades-STA

Different success cases of our algorithm on the Charades-STA dataset can be seen in Figure 1 and Figure 2. It is interesting to see that as soon as our method can attend frames inside of the action the localization layer can predict a good start and end temporal location.

Failure cases of our method are presented in Figure 3 and Figure 4. We can see that attention layer gets confused in the first example, it does not know what is the most important feature for the query, making the localization layer fail to predict good temporal localization. Figure 4 shows that the attention layer gets confused with frames that has similar appearance, Figure 5.

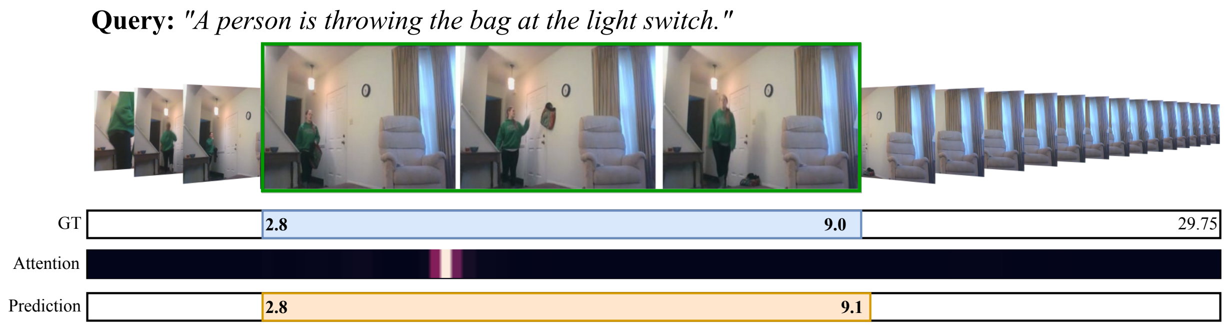

Figure 1: Success case of our method on Charades-STA dataset.

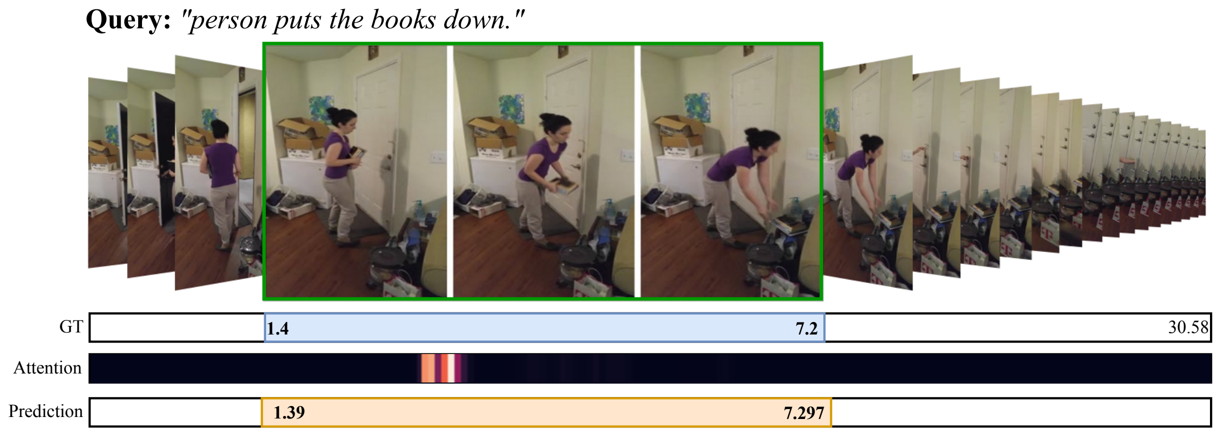

Figure 2: Success case of our method on Charades-STA dataset.

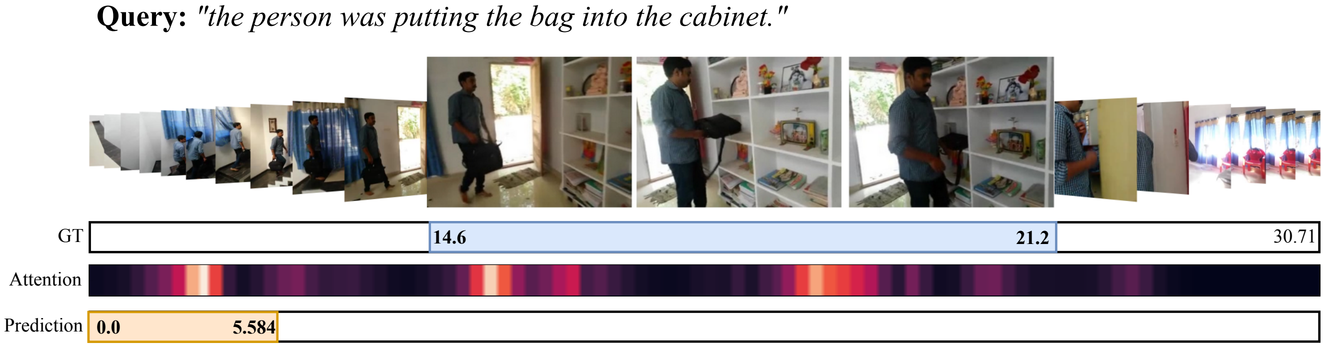

Figure 3: Failure case of our method on Charades-STA

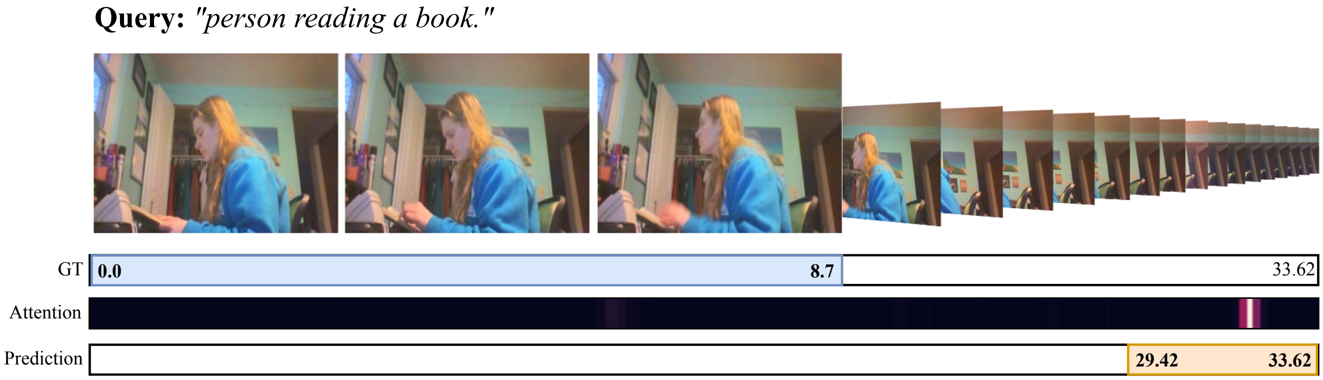

Figure 4: Failure case of our method on Charades-STA

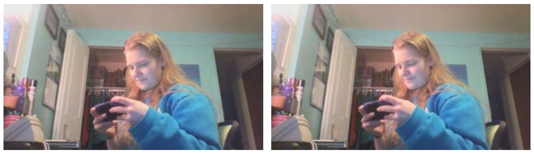

Figure 5: Confusing frames

2. Activity-Net Caption

Although videos in ActivityNet Caption are much longer than videos in Charades-STA, our method still can get good localization performance if the attention layer makes a good job, as can be seen in Figure 6 and Figure 7. Notice that Figure 7 shows a long action that spans more than 2.5 minutes.

Failure cases of our method on ActivityNet Caption dataset are presented in Figure 8 and Figure 9. Our method has similar difficulties in Charades-STA and ActivityNet Caption. Every time that the attention fails to focus in frames inside of the corresponding moment the localization layer cannot predict the correct temporal localization of the query. Figure 10 shows frames that are also related to query in Figure 9. These image suggest that our method can understand what a credits is and where is located but cannot distinguish *ending* or *starting*

Query: *"We then see one man climbing a sheer cliff."*

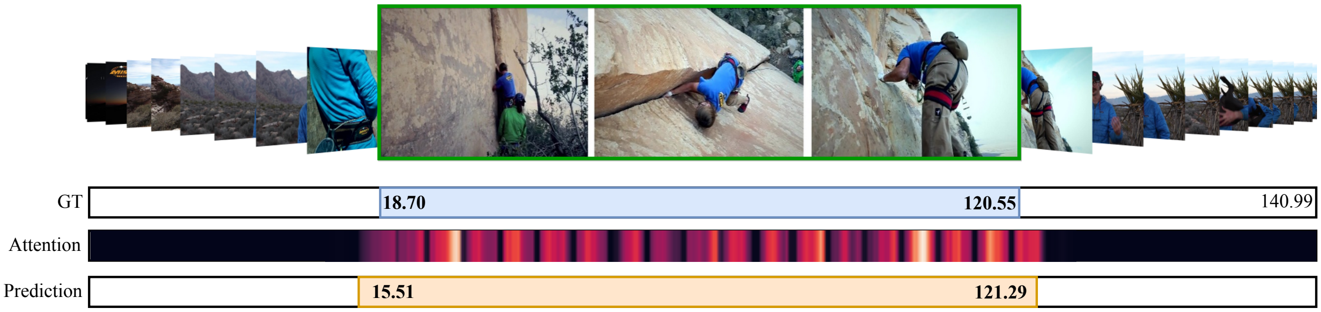

Figure 6: Success case of our method in ActivityNet Caption dataset.

Query: *"They then get up with jump ropes and the two begin doing various types of jumps."*

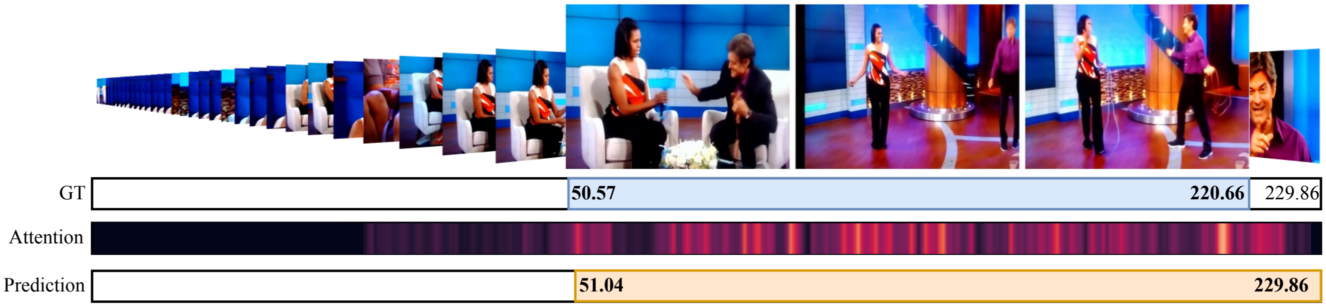

Figure 7: Success case of our method in ActivityNet Caption dataset.

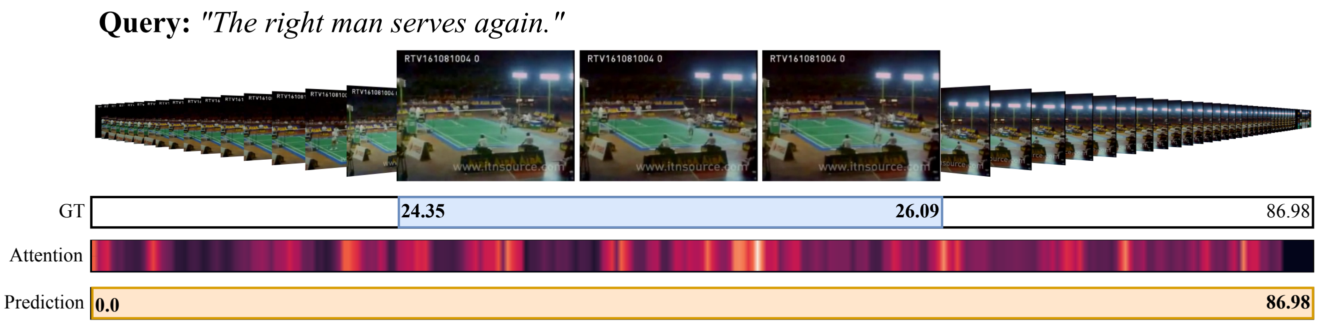

Figure 8: Failure case of our method in ActivityNet Caption dataset.

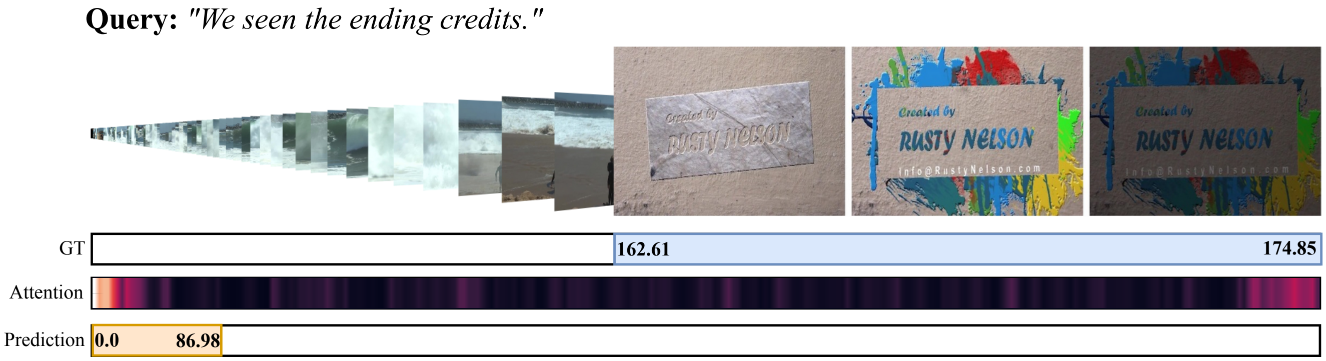

Figure 9: Failure case of our method in ActivityNet Caption dataset.

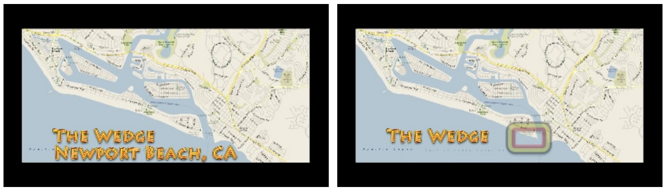

Figure 10: Confusing frames.