Co-Grounding Networks with Semantic Attention for Referring Expression Comprehension in Videos (Supplementary Material)

Sijie Song[1], Xudong Lin[2], Jiaying Liu[1], Zongming Guo[1], Shih-Fu Chang[2]
[1] Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[2] DVMM Lab, Columbia University, New York, NY, USA

In this supplementary, we provide more implementation details of co-grounding feature learning in Sec. 1. More overall grounding results are presented in Sec. 2 to show the effectiveness of semantic attention and co-grounding feature learning. In Sec. 3, by visualizing attention patterns, we analyze how semantic attention improves the single-frame accuracy and how co-grounding feature learning improves grounding consistency across frames.

1. Implementation details of co-grounding feature learning

As the overall framework shown in the main paper (Figure 2), during training, we need to sample two frames from a video as input, and then forward the network for co-grounding feature learning. In practice, we take two adjacent frames as a training pair. During inference, we consider multiple neighboring T_r frames as reference. When we predict the bounding box for the t^*-th frame, given the visual feature $\mathbf{F}_{t^*} \in \mathbb{R}^{H\times W\times D}$ from the visual encoder and its neighboring features $\mathcal{F} = [\mathbf{F}_{t^*-\Delta t}, .., \mathbf{F}_{t^*+\Delta t}]$, where $\Delta t = \lfloor T_r/2 \rfloor$, we calculate the normalized affinity matrix $M \in \mathbb{R}^{HW\times HWT_r}$. Then we integrate $\mathcal{F}$ with the matrix M as illustrated in Sec.3.2 in the main paper, and obtain $\mathbf{F}'_{t^*}$. The final visual feature is generated by $\mathbf{V}_{t^*} = \mathbf{F}'_{t^*} \oplus \mathbf{F}_{t^*}$, where $\oplus$ is concatenation along the channel. Table 1 shows the effect of T_r. We set $T_r = 5$ in our experiments.

Table 1. The effect of T_r on the VID dataset in terms of Acc.@0.5 and mIoU.

	T_r=3	T_r=5	T_r=7
Acc.@0.5	59.40	59.48	59.51
mIoU	0.490	0.494	0.495

2. Overall grounding results

More grounding results are presented in Figure 1. Each row shows the results obtained from different models. The ground-truth bounding boxes are denoted in green, while the predicted results are denoted in blue. The baseline model is from [1]. With semantic attention, SL-Att. is able to improve the single-frame grounding accuracy compared to baseline. However, bounding box drifting is observed (see the second rows in Figure 1(a), Figure 1(b) and Figure 1(c)). It is mainly because the features from a single frame are vulnerable to scene dynamics. With co-grounding feature learning, the visual features are enhanced by neighboring features, leading to more robust representations. Therefore, we obtain stable and consistent results from CG-SL-Att. (see the third rows in Figure 1(a), Figure 1(b) and Figure 1(c)).

3. Visualizations on semantic attention

• **How does semantic attention improve the single-frame accuracy?** We provide more visualizations on semantic attention obtained from SL-Att. in Figure 2 and Figure 3, to see how the model improves grounding accuracy for a single frame compared to baseline [1]. The expressions are shown above the images. In Figure 2(a) and Figure 2(b), we observe similar subject attention patterns for the visual features, corresponding to '*smallest elephant*' and '*gray elephant*' in the expressions, respectively. However, the location words '*left*' and '*right*' lead to different location attention maps. The cues in location help the model predict the correct bounding boxes (shown in red in the third columns). The other example shown in Figure 3 illustrates the cues for subject help distinguish the grounding results. In subject attention, the words related to '*ship floats*', '*padding its tail*' are attended, and subject attention maps show different response according to the expression. Though there is no dominant word indicating location in the expressions of Figure 3(a) and Figure 3(b), we still obtain reasonable location attention maps, which are generated under the guidance of subjects.

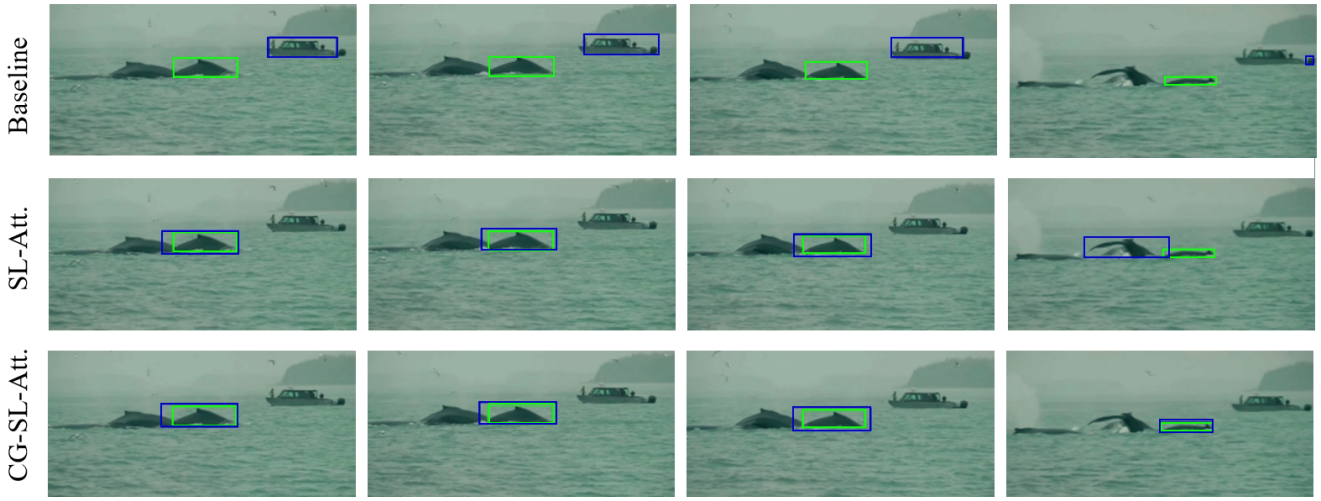
• **How does co-grounding feature learning improve grounding consistency across frames?** Figure 4 and Figure 5 show the attention comparisons between SL-Att. and

CG-SL-Att. to illustrate how co-grounding features improve grounding consistency. For the same expression, the attentions from the two models share similar patterns. However, different response for visual features are observed, and further leads to different predictions. In Figure 4(a), SL-Att. predicts the bounding box incorrectly due to the wrong subject response (indicated by the black arrow). With co-grounding feature learning, such subject response is eliminated as shown in Figure 4(b), providing correct cues for CG-SL-Att. to predict the correct bounding box. The similar phenomenon is also observed in Figure 5. Without reference from neighboring frames, SL-Att. regards the stone on the left as '*antelope*' (indicated by the black arrow in Figure 5(a)) in the subject attention map. On the contrary, co-grounding features are more robust for cross-modal matching and we see clear response for '*antelope*' in the subject attention map of Figure 5(b). Besides, correct subject response provides correct guidance for reasonable location attention maps (see the second columns in Figure 4 and Figure 5). The final predicted results (in blue) are shown in the third columns, where ground-truths are in green. We see that CG-SL-Att. outperforms SL-Att. due to correct referring cues obtained from robust visual features.

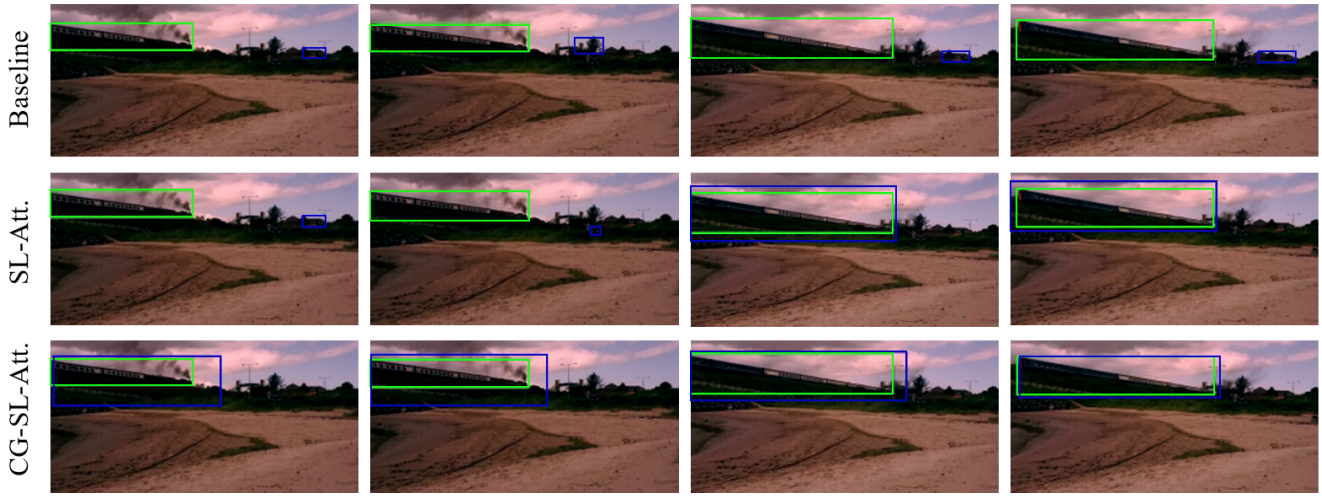
References

[1] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Int. Conf. Comput. Vis.*, pages 4683–4693, 2019. 1, 3, 4

(a) The smallest elephant on the left in the water paces and drinks in the water.



(b) The black whale on the right is jumping down the sea, making a lot of water spray.



(c) A train with lots of carriages is moving from left to right crossing a bridge with black smoke.

Figure 1. Overall grounding results. The expression queries are in sub-captions. The ground-truths are in green, and the predicted results are in blue. The baseline results are obtained by per-frame inference with [1]. It is observed that SL-Att. is able to improve grounding accuracy compared to baseline. CG-SL-Att. further improves grounding consistency across frames.

	The	smallest	elephant	on	the	left	in	the	water	paces	and	drinks	in	the	water
sub. att.	0.00	0.50	0.09	0.00	0.00	0.04	0.00	0.00	0.16	0.12	0.00	0.09	0.00	0.00	0.06
loc. att.	0.00	0.05	0.01	0.03	0.01	0.69	0.01	0.00	0.04	0.05	0.05	0.04	0.01	0.00	0.02

sub. att. map loc. att. map results

(a) The subject attention map shows response to '*smallest elephant*', while the location attention map shows response to '*left*'.

	The	gray	elephant	is	walking	alone	on	right	side	of	the	river	near	a	tree
sub. att.	0.00	0.16	0.02	0.20	0.01	0.12	0.01	0.00	0.00	0.05	0.12	0.02	0.00	0.21	0.08
loc. att.	0.01	0.00	0.04	0.00	0.25	0.00	0.07	0.61	0.02	0.00	0.00	0.02	0.01	0.00	0.01

sub. att. map loc. att. map results

(b) The subject attention map shows response to '*gray elephant*', while the location attention map shows response to '*right*'.

Figure 2. Attention visualization for SL-Att. to show semantic attention provides explicit referring cues. The bounding boxes of ground-truths, baseline and SL-Att. are in green, blue and red, respectively. The baseline results are obtained by per-frame inference with [1].

	A	ship	in	the	distance	floats	on	the	right	hand	size
sub. att.	0.00	0.18	0.00	0.00	0.15	0.22	0.00	0.00	0.08	0.27	0.09
loc. att.	0.21	0.01	0.15	0.05	0.01	0.01	0.57	0.04	0.01	0.01	0.01

sub. att. map loc. att. map results

(a) The subject attention map shows response to '*ship floats*', and the location attention map shows response to the subject accordingly.

	A	black	whale	is	patting	its	tail	out	of	the	water
sub. att.	0.00	0.05	0.08	0.01	0.41	0.16	0.17	0.04	0.00	0.00	0.07
loc. att.	0.32	0.02	0.02	0.45	0.01	0.03	0.01	0.01	0.04	0.10	0.02

sub. att. map loc. att. map results

(b) The subject attention map shows response to '*padding its tail*', and the location attention map shows response to the subject accordingly.

Figure 3. Attention visualization for SL-Att. to show semantic attention provides explicit referring cues. The bounding boxes of ground-truths, baseline and SL-Att. are in green, blue and red, respectively. The baseline results are obtained by per-frame inference with [1].

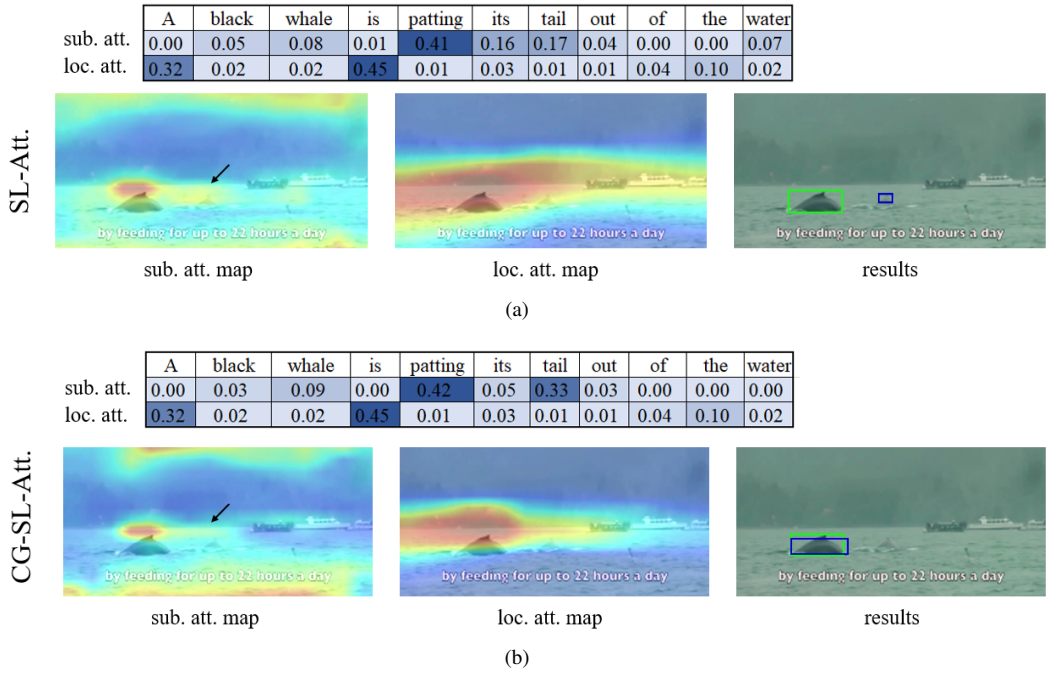


Figure 4. Attention comparison between SL-Att. and CG-SL-Att.. It is observed the incorrect subject response from SL-Att. (black arrow in (a)) is eliminated by co-grounding feature learning (black arrow in (b)), providing CG-SL-Att. with correct subject cues. We show the final predicted bounding boxes in the third columns. The ground-truths are in green, and the predicted results are in blue.

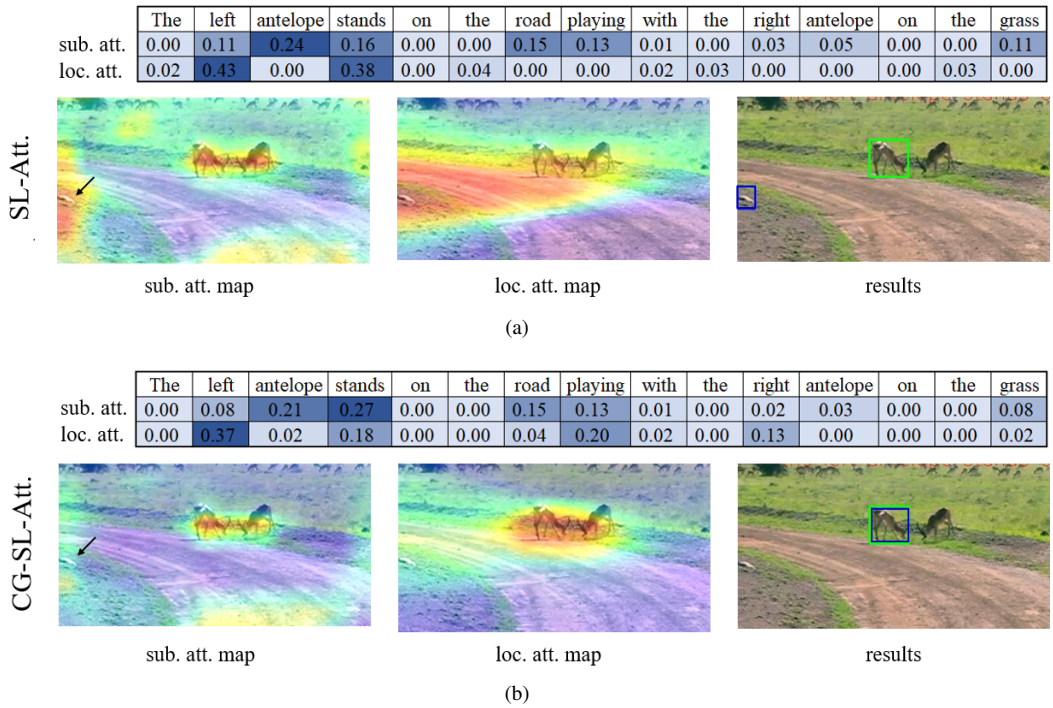


Figure 5. Attention comparison between SL-Att. and CG-SL-Att.. In (a), SL-Att. regards the stone as '*antelope*' (black arrow in (a)) in the subject attention map. In (b), the subject attention map is more accurate due to co-grounding feature learning. Besides, subject response also provides guidance for location attention maps. We show the final predicted bounding boxes in the third columns. The ground-truths are in green, and the predicted results are in blue.