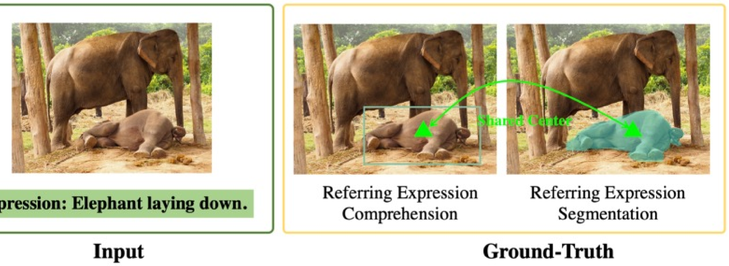
LoA-Trans: Enhancing Visual Grounding by Location-Aware Transformers

Ziling Huang[1,2], Shin'ichi Satoh[2,1]
[1]The University of Tokyo [2]National Informatics Institute

Constraining REC and RES for Identical Object



Problems:

- Previous work does not explicitly assume that RIC and RIS address the same object
- Previous work lacks any specific coordination between tasks.

Our contributions:

- A center prompt to ensure REC and RES reference the same object.
- A query selection mechanism for initializing location-aware queries
- A TaskSyn Network to facilitate effective information exchange.

REC Leaderboard(Pr@50)

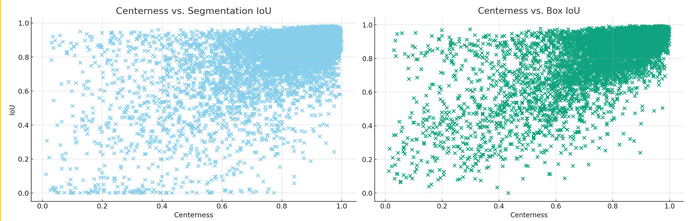
Model	Refcoco (val)	Refcoco (testA)	Refcoco (testB)	Refcoco+ (val)	Refcoco+ (testA)	Refcoco+ (testB)	Refcocog (val)	Refcocog (test)
TransVG	80.83	83.38	76.94	68.00	72.46	59.24	68.71	67.98
QRNet	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03
RefTR	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40
SeqTR	81.23	85.00	76.08	68.82	75.37	58.78	71.35	71.58
VG-LAW	86.62	89.32	83.16	76.37	81.04	67.50	76.90	76.96
LoA-Trans-S	87.59	90.17	82.98	78.68	83.93	69.83	79.58	79.29
LoA-Trans-B	**87.75**	**90.60**	**84.81**	**79.56**	**84.95**	**71.75**	**80.80**	**80.18**

RES Leaderboard(mIoU)

Model	Refcoco (val)	Refcoco (testA)	Refcoco (testB)	Refcoco+ (val)	Refcoco+ (testA)	Refcoco+ (testB)	Refcocog (val)	Refcocog (test)
VLT	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
LAVT	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
ReLA	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
RefTR	70.56	73.49	66.57	61.08	64.69	52.73	58.73	58.51
SeqTR	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
VG-LAW	75.62	77.51	72.89	66.63	70.38	58.89	65.63	66.08
LoA-Trans-S	76.03	77.90	72.57	67.85	72.21	60.29	67.44	67.97
LoA-Trans-B	**76.66**	**78.60**	**74.17**	**69.40**	**73.59**	**62.90**	**69.01**	**68.77**

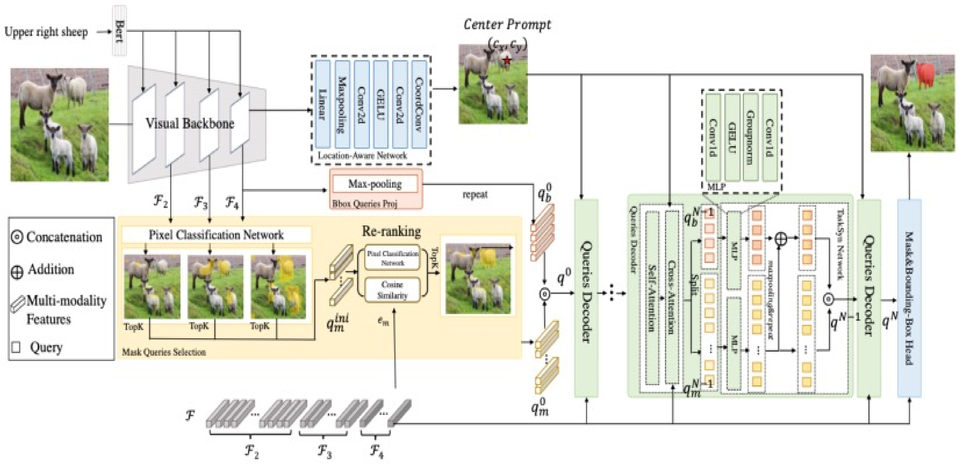
Center prompt vs. Estimated Results

$$Centerness = \sqrt{\frac{\min(l^*, r^*)\,\min(t^*, b^*)}{\max(l^*, r^*)\,\max(t^*, b^*)}}$$



- The higher centerness, the higher IoU score.
- Performance improves notably at mid to high centerness ranges
- In ranges ([0.7, 1.0)), the model achieves near-perfect accuracy

LoA-Trans Framework



Visualization

Ablation Study

#	Methods	RIS(mIoU)	RIC(Pr@50)
1	w/o TaskSyn Network	**74.58**	**86.52**
2	Random queries	**75.67**	**87.06**
3	No Selection	**Out of Memory**	
4	Ours	**76.03**	**87.59**

Acknowledge

This work is partly supported by JSPS KAKENHI Grant Number JP23K24876 and JST ASPIRE Program Grant Number JPMJAP2303.