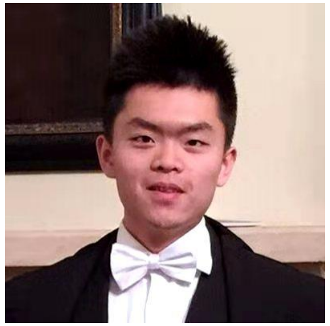
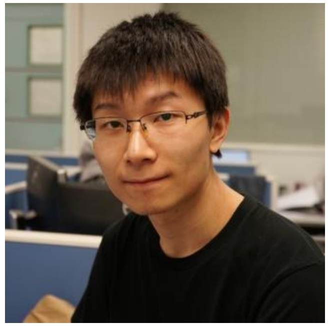
Microsoft

LAVT: Language-Aware Vision Transformer for Referring Image Segmentation



Zhao Yang[1*] Jiaqi Wang[2*] **Yansong Tang**[5,1#] Kai Chen[2,4] Hengshuang Zhao[3,1] Philip H.S. Torr[1]

(*Equal Contribution, #Corresponding Author)

[1]University of Oxford, [2]Shanghai AI Laboratory, [3]The University of Hong Kong,
[4]SenseTime Research, [5]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

https://github.com/yz93/LAVT-RIS

Background

Background

Instance Segmentation

HTC [Chen et al. CVPR2019]

Fixed Category

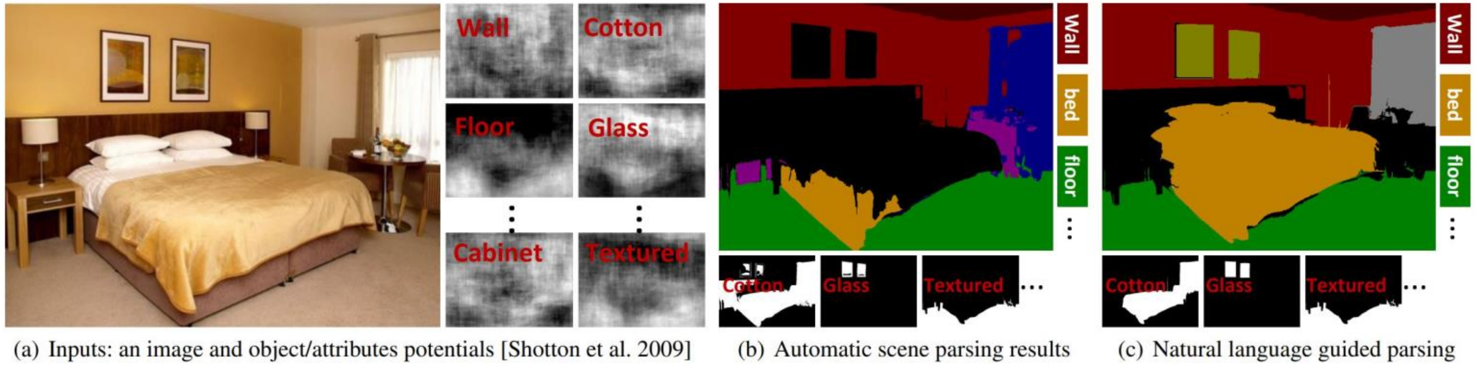
"man on the left"

Referring Segmentation

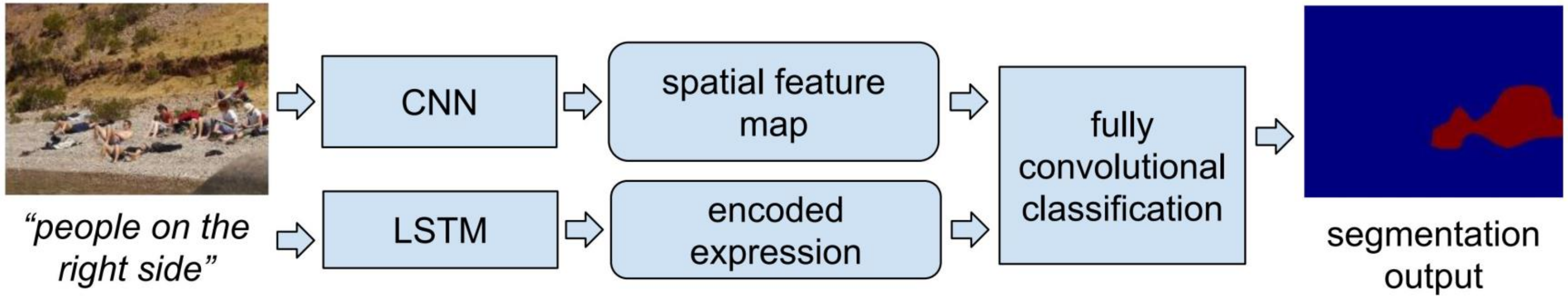
LAVT (Ours)

Open Vocabulary

Related Work



ImageSpirit: Verbal Guided Image Parsing [Cheng et al. TOG2014]

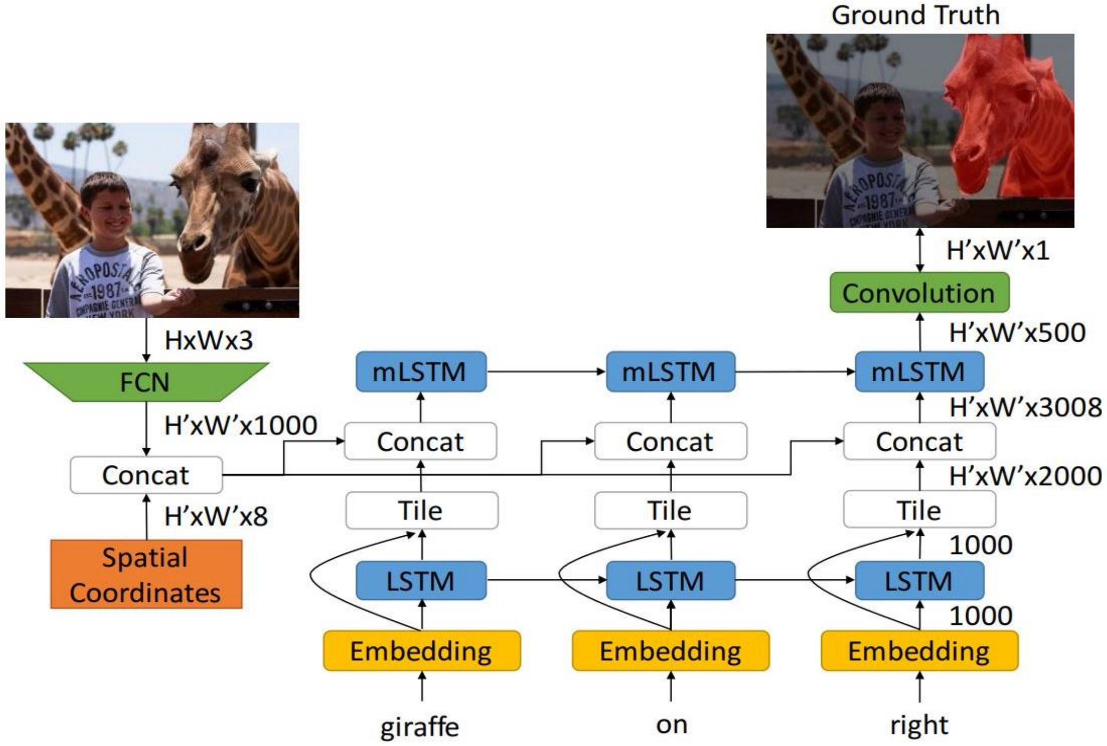


CNN+LSTM [Hu et al. ECCV2016]

Related Work

CNN+LSTM
[Hu et al. ECCV2016]

RMI
[Liu et al. ICCV2017]

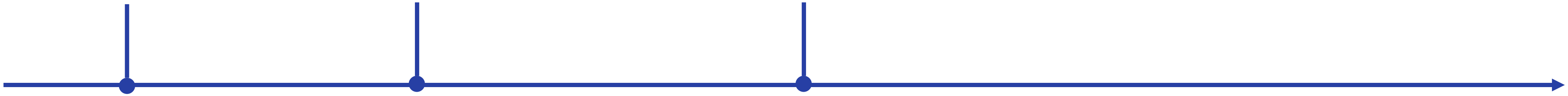


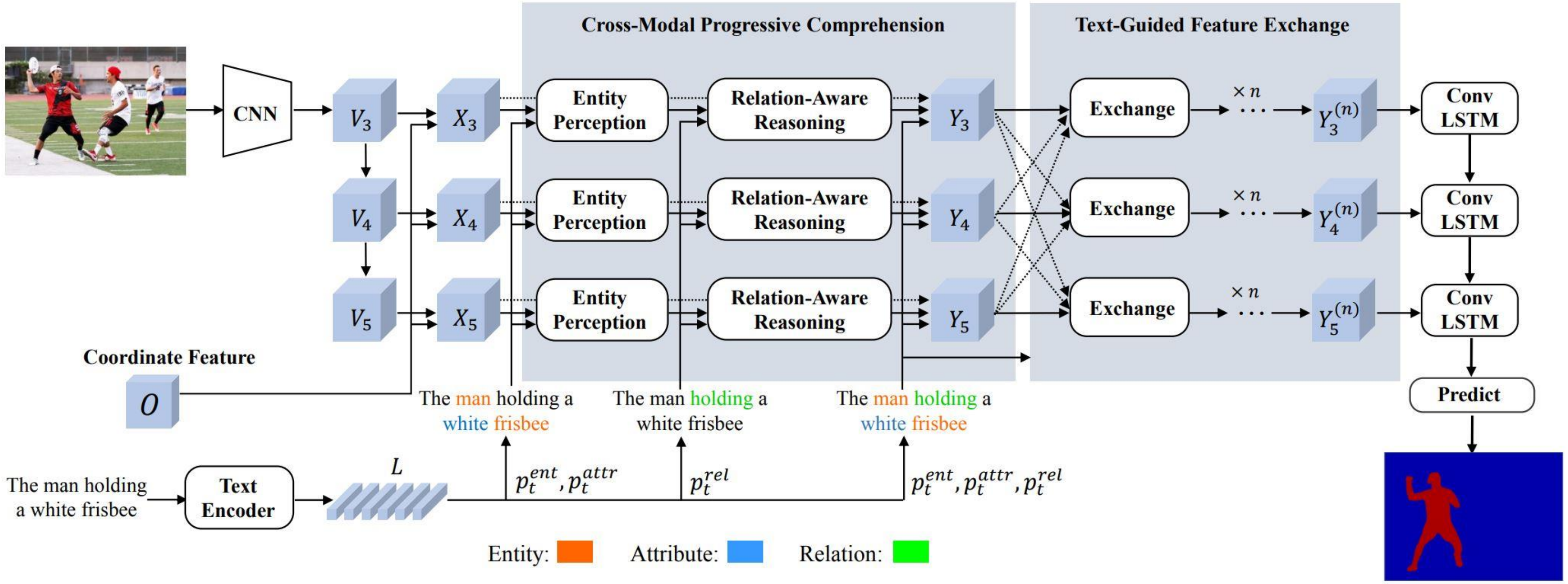
Related Work

CNN+LSTM
[Hu et al. ECCV2016]

RMI
[Liu et al. ICCV2017]

CMPC
[Huang et al. CVPR2020]





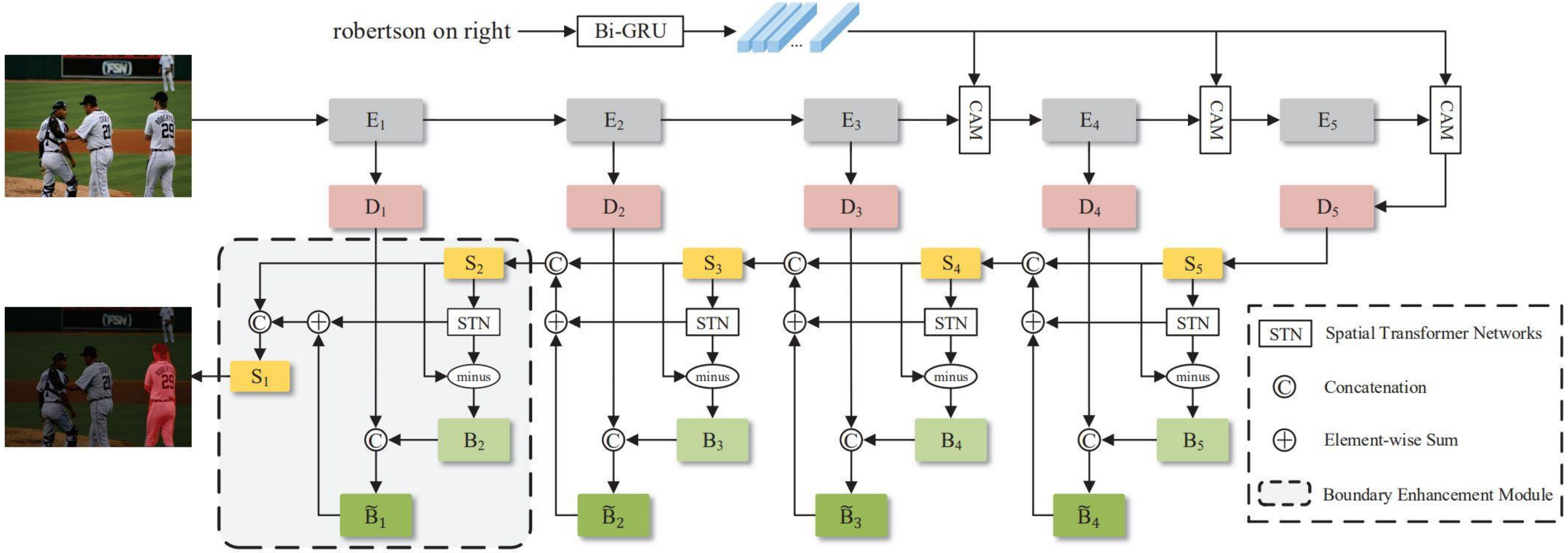
Related Work

CNN+LSTM
[Hu et al. ECCV2016]

RMI
[Liu et al. ICCV2017]

CMPC
[Huang et al. CVPR2020]

EFN
[Feng et al. CVPR2021]



Related Work

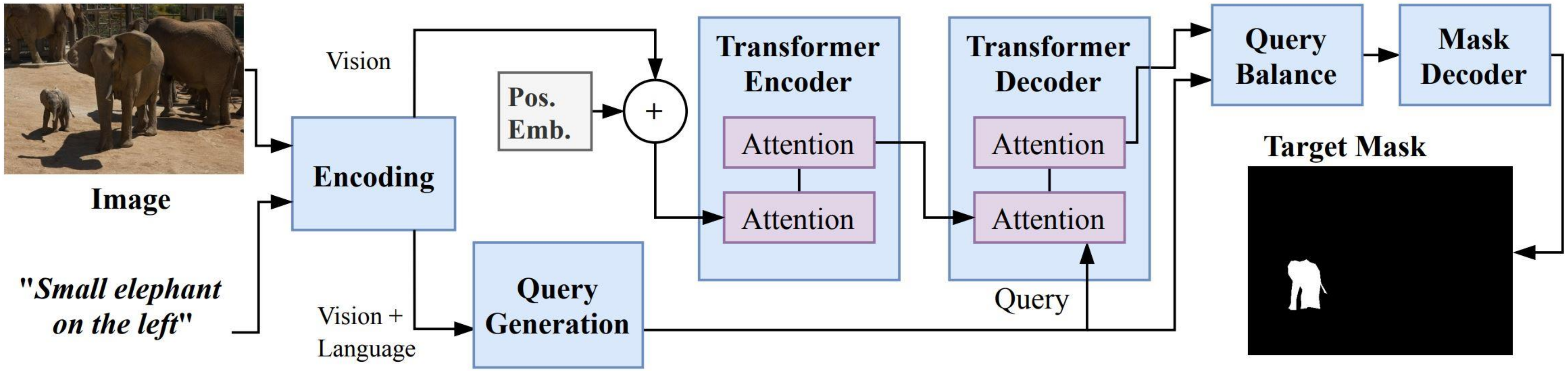
CNN+LSTM
[Hu et al. ECCV2016]

RMI
[Liu et al. ICCV2017]

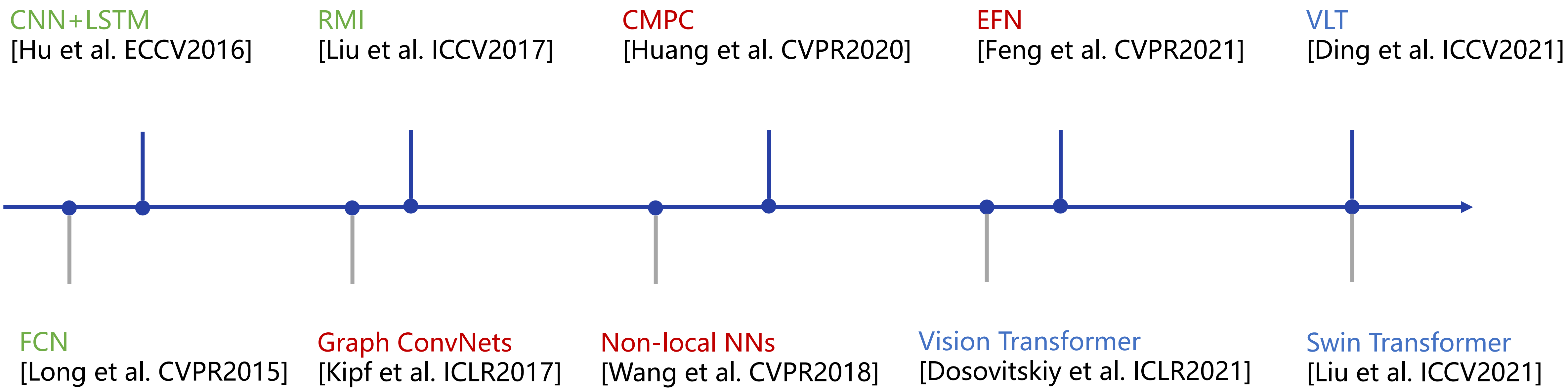
CMPC
[Huang et al. CVPR2020]

EFN
[Feng et al. CVPR2021]

VLT
[Ding et al. ICCV2021]



Related Work

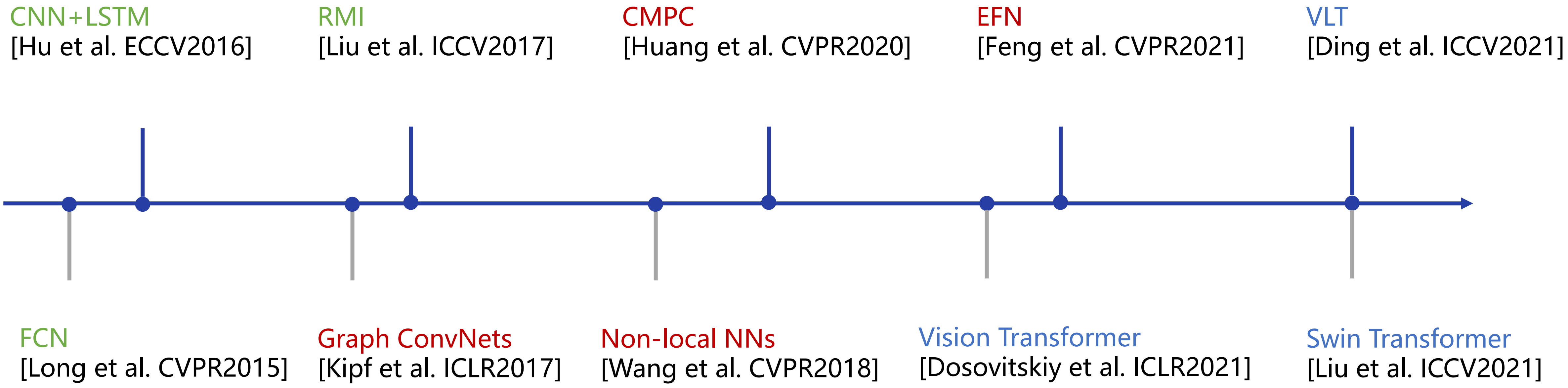


A shift of paradigm is happening for RIS on two dimensions:

1. Specialized modeling of cross-modal correspondences - Transformers

2. Convergence on the vision and language backbones - Transformers

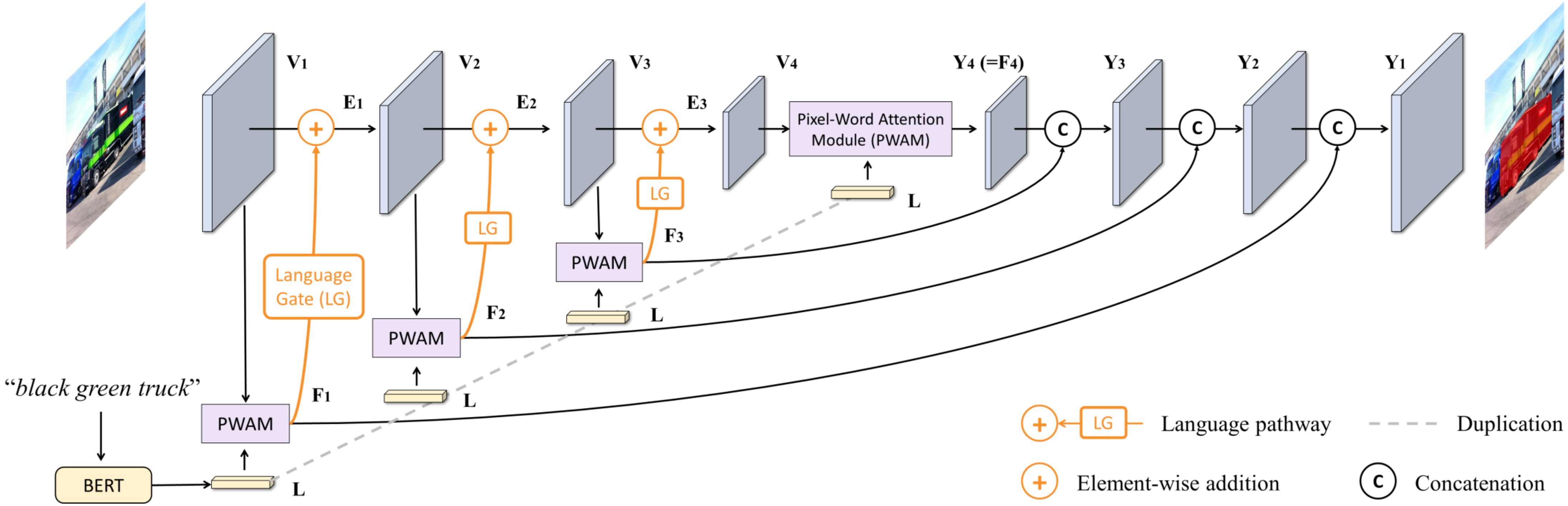
Related Work



Transformers have shown to be the best on both fronts

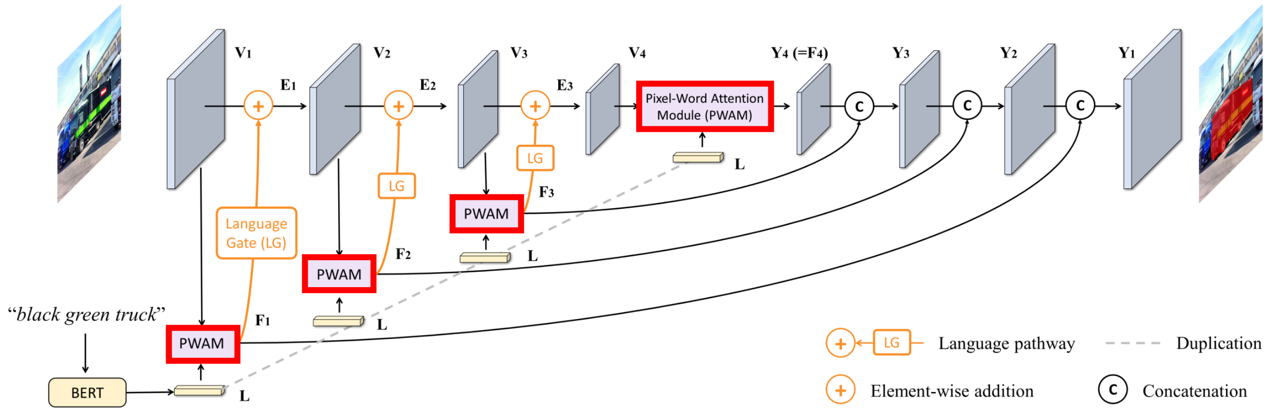
Why not unite efforts to build a more unified approach: let cross-modal correspondence modeling benefit from the vision Transformer encoder?

Approach

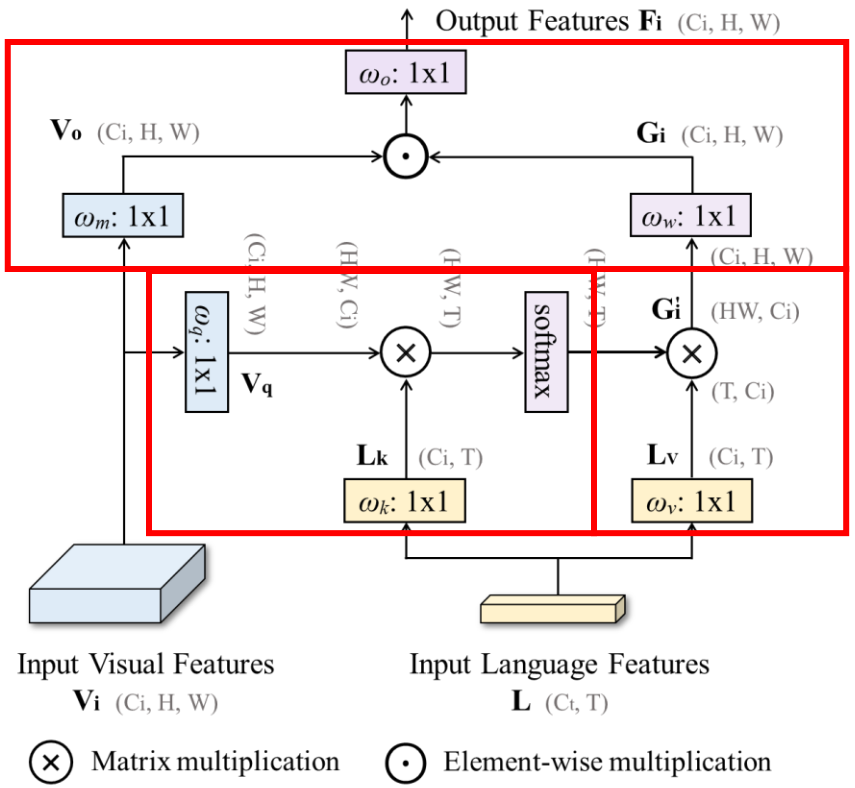


V1
V2
V3
V4
Y4 (=F4)
Y3
Y2
Y1
E1
E2
E3
Pixel-Word Attention Module (PWAM)
L
LG
F3
PWAM
Language Gate (LG)
F2
PWAM
F1
"black green truck"
PWAM
BERT
Language pathway
Duplication
Element-wise addition
Concatenation

Approach

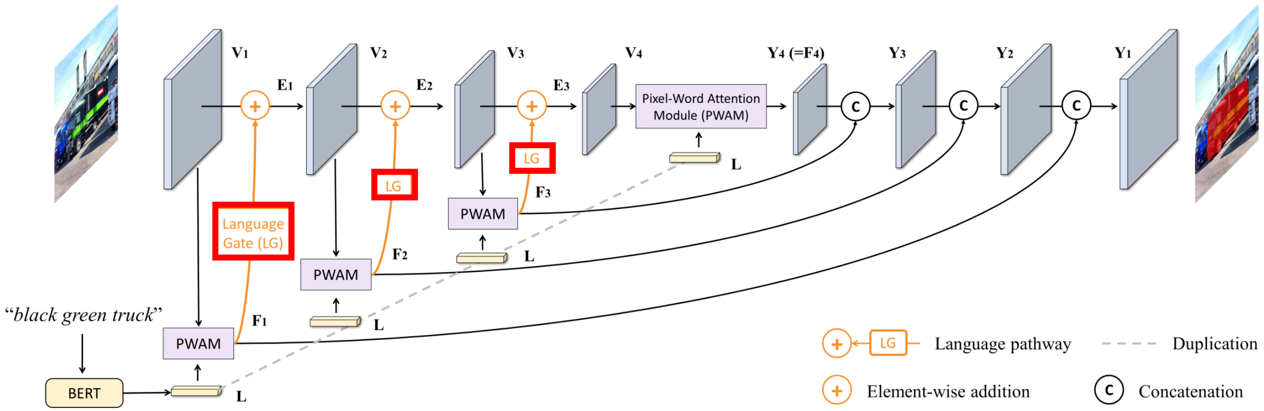


Framework

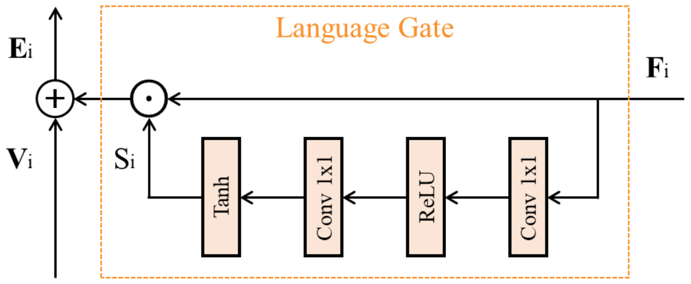


Pixel-Word Attention Module (PWAM)

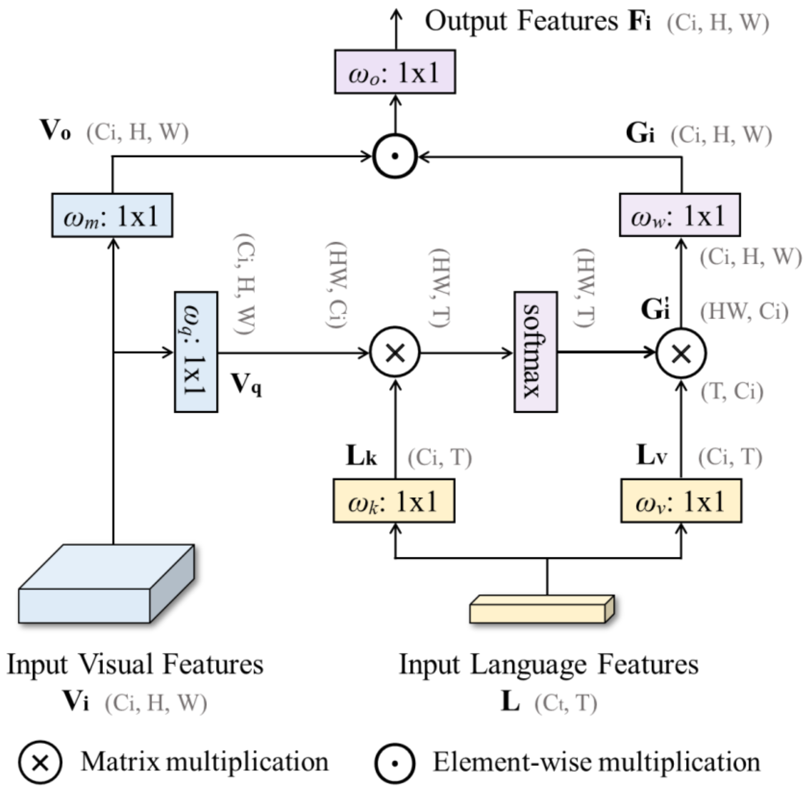
Approach



Framework



Language Path (LP)



Pixel-Word Attention Module (PWAM)

Experiment

Dataset

	RefCOCO	RefCOCO+	G-Ref
#Images	19,994	19,992	26,711
#Objects	50,000	49,856	54,822
#Expressions	142,209	141,564	104,560

Evaluation Metric

- Overall Intersection-over-union (oIoU)
- Mean Intersection-over-union (mIoU)
- Precision at the α threshold values (P@α)

RefCOCO, RefCOCO+[Yu et al. ECCV2016]; G-Ref [Mao et al. CVPR2016]

Experiment – Comparison with State-of-the-art Methods

Method	Language Model	RefCOCO			RefCOCO+			G-Ref		
		val	test A	test B	val	test A	test B	val (U)	test (U)	val (G)
DMN [43]	SRU	49.78	54.83	45.13	38.88	44.22	32.29	-	-	36.76
RRN [30]	LSTM	55.33	57.26	53.93	39.75	42.15	36.11	-	-	36.45
MAttNet [63]	Bi-LSTM	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-
CMSA [62]	None	58.32	60.61	55.09	43.76	47.60	37.89	-	-	39.98
CAC [8]	Bi-LSTM	58.90	61.77	53.81	-	-	-	46.37	46.95	44.32
STEP [5]	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-	46.40
BRINet [23]	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-	48.04
CMPC [24]	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-	49.05
LSCM [25]	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-	48.05
CMPC+ [34]	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-	49.89
MCN [41]	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
EFN [15]	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-	51.93
BUSNet [58]	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56
CGAN [40]	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
LTS [27]	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
VLT [13]	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76
LAVT (Ours)	BERT	**72.73**	**75.82**	**68.79**	**62.14**	**68.38**	**55.10**	**61.24**	**62.09**	**60.50**

Table 1. Comparison with state-of-the-art methods in terms of overall IoU on three benchmark datasets. U: The UMD partition. G: The Google partition. We refer to the language model of each reference method as the main learnable function that transforms word embeddings before multi-modal feature fusion. Interested readers can refer to the respective papers for embedding initialization and other details.

Experiment – Ablation Study

LP	PWAM	P@0.5	P@0.7	P@0.9	oIoU	mIoU
✓	✓	**84.46**	**75.28**	**34.30**	**72.73**	**74.46**
	✓	81.46	70.80	30.95	70.78	71.96
✓		81.76	72.76	32.46	71.03	72.31
		77.87	66.93	27.95	68.82	68.87

Table 2. Main ablation results on the RefCOCO validation set.

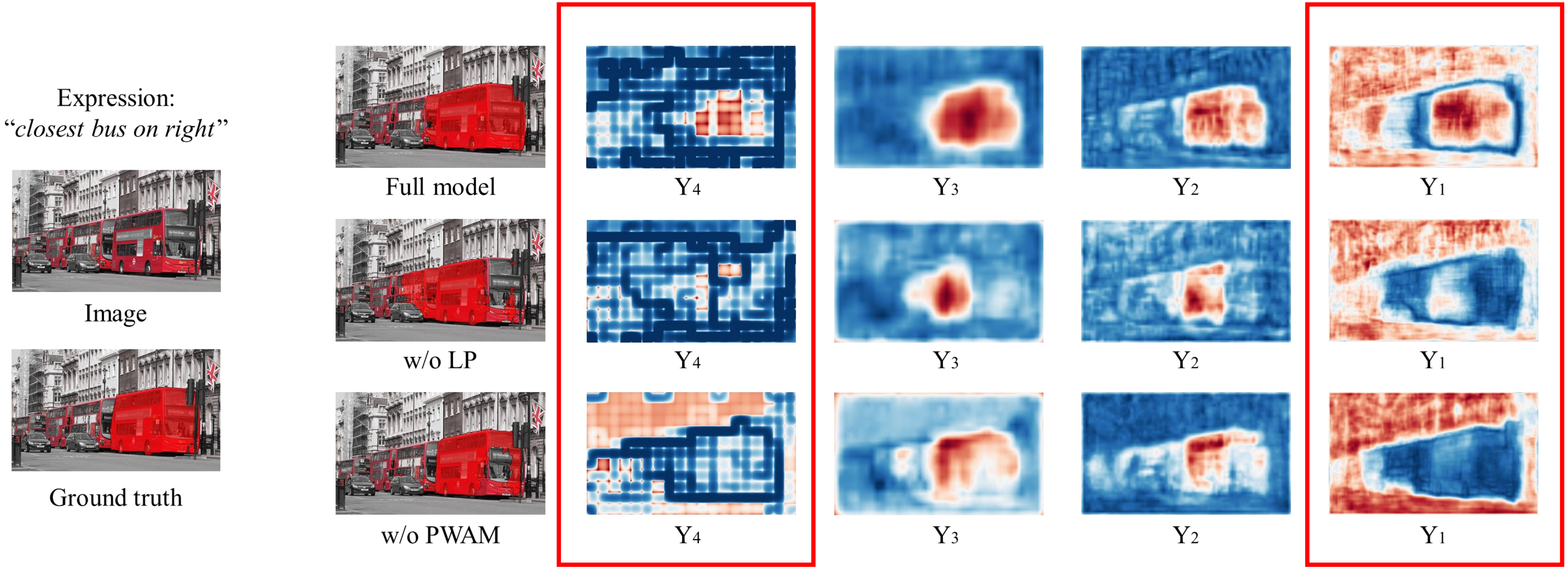
Method	P@0.5	P@0.7	P@0.9	oIoU	mIoU
LTS (Swin-B+BERT) [27]	80.59	69.48	26.13	69.94	70.56
EFN (Swin-B+BERT) [15]	82.55	73.27	31.68	70.76	72.95
VLT (Swin-B+BERT) [13]	83.24	72.81	24.64	70.89	71.98
Ours + VLT [13]	**84.57**	75.14	26.36	72.12	73.57
Ours	84.46	**75.28**	**34.30**	**72.73**	**74.46**

Table 4. Comparison between our method, LTS [27], VLT [13], and EFN [15] on the RefCOCO validation set, where all models use the same backbone, language model, and training recipes.

	P@0.5	P@0.7	P@0.9	oIoU	mIoU
(a) activation function in the language gate (LG)					
Tanh (*)	**84.46**	**75.28**	**34.30**	**72.73**	**74.46**
Sigmoid	81.89	72.71	33.35	70.49	72.47
(b) normalization layer in pixel-word attention module (PWAM)					
InstanceNorm (*)	**84.46**	**75.28**	**34.30**	**72.73**	**74.46**
LayerNorm	82.97	74.15	33.99	71.92	73.32
BatchNorm	82.89	73.82	33.53	71.59	73.09
None	81.91	72.73	33.11	70.66	72.34
(c) features used for final classification					
F_4, F_3, F_2, F_1 (G*)	**84.46**	**75.28**	34.30	**72.73**	**74.46**
F_4, F_3, F_2, F_1 (NG)	84.00	74.96	33.47	72.24	73.94
E_4, E_3, E_2, E_1 (G)	83.84	74.96	34.48	72.06	73.98
E_4, E_3, E_2, E_1 (NG)	84.33	74.94	**34.77**	72.27	74.12
V_4, V_3, V_2 (G)	83.36	74.47	32.61	71.38	73.29
V_4, V_3, V_2 (NG)	83.83	74.76	32.14	72.29	73.67
(d) multi-modal attention module					
PWAM (*)	**84.46**	**75.28**	**34.30**	**72.73**	**74.46**
BCAM [23]	82.26	72.81	33.31	70.19	72.42
GA (GARAN) [40, 41]	83.22	74.09	32.71	71.20	73.16

Table 3. Ablation studies on the RefCOCO validation set. (G) indicates that LG is adopted in the language pathway and (NG) indicates the opposite. Rows with (*) indicate default choices.

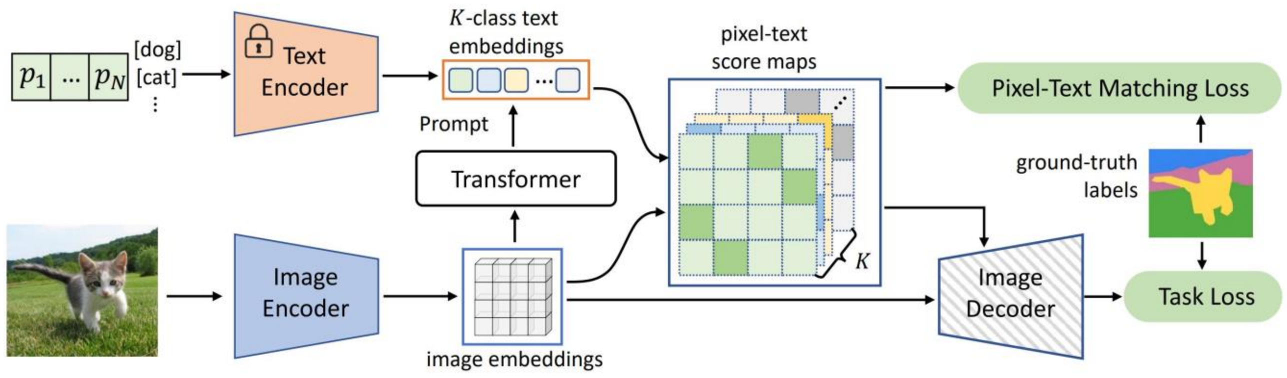
Experiment – Visualized Results



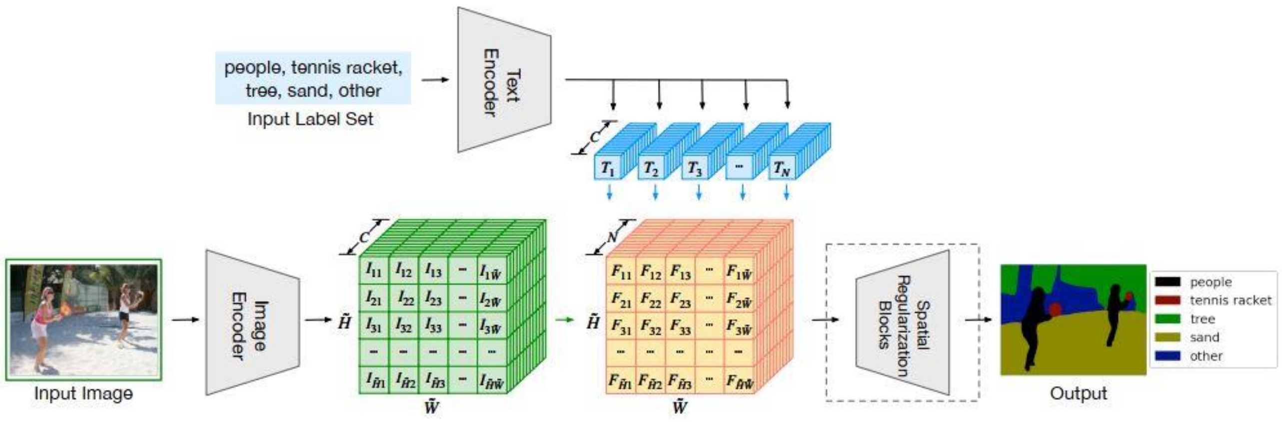
Conclusion

- LAVT: leveraging the multi-stage design of a vision Transformer for jointly encoding multi-modal inputs
- Experimental results on three benchmarks have demonstrated its advantage with respect to the state of the art
- Code available at https://github.com/yz93/LAVT-RIS

Future Work – Language-Guided Dense Prediction



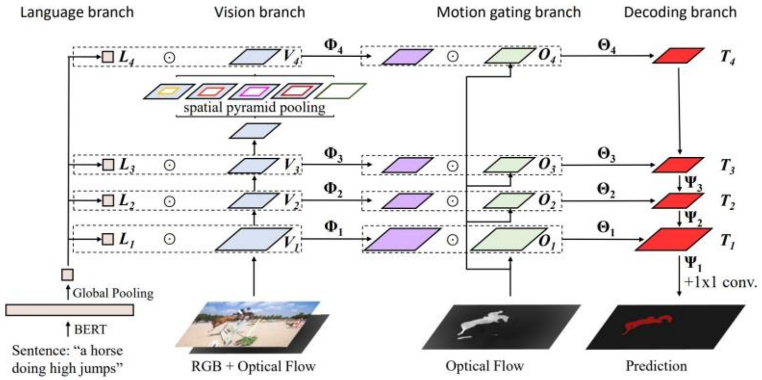
DenseCLIP
[Rao*, Zhao* et al. CVPR2022]



Language-Driven
Semantic Segmentation
[Li et al. ICLR2022]

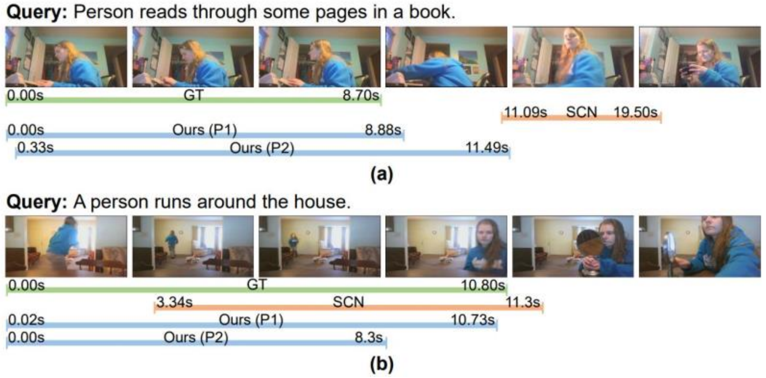
Future Work – Referring Segmentation in Other Fields

Referring Video Segmentation



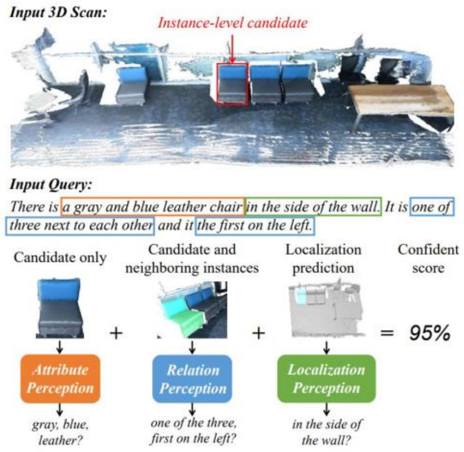
HINet
[Yang*, Tang* et al. BMVC2021]

Temporal Sentence Grounding



CPL
[Zheng et al. CVPR2022]

Referring 3D Instance Segmentation



InstanceRefer
[Yuan et al. ICCV2021]

Thanks All!