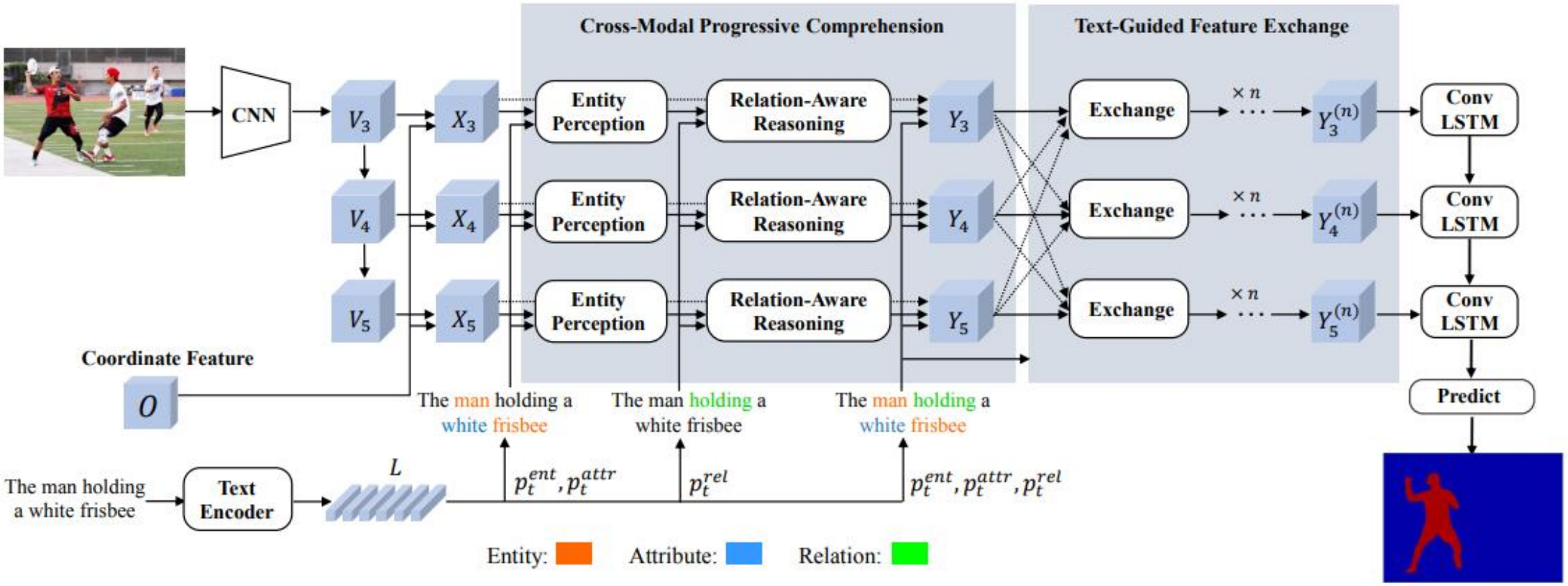
LAVT: Language-Aware Vision Transformer for Referring Image Segmentation

Zhao Yang[1], Jiaqi Wang[2], Yansong Tang[1], Kai Chen[2,3], Hengshuang Zhao[1,4], Philip H.S. Torr[1]
[1]University of Oxford, [2]Shanghai AI Laboratory,
[3]SenseTime Research, [4]The University of Hong Kong

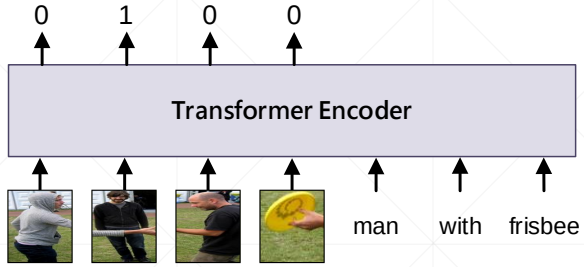
Mengxue

Review

- **CNN-based Method**



Review



- **Transformer-based Method**
- 1) Pre-processed with trained-detector: Mask-rcnn→get segmentation from box directly
- 2) VLT: Vision-Language Transformer and Query Generation for Referring Segmentation

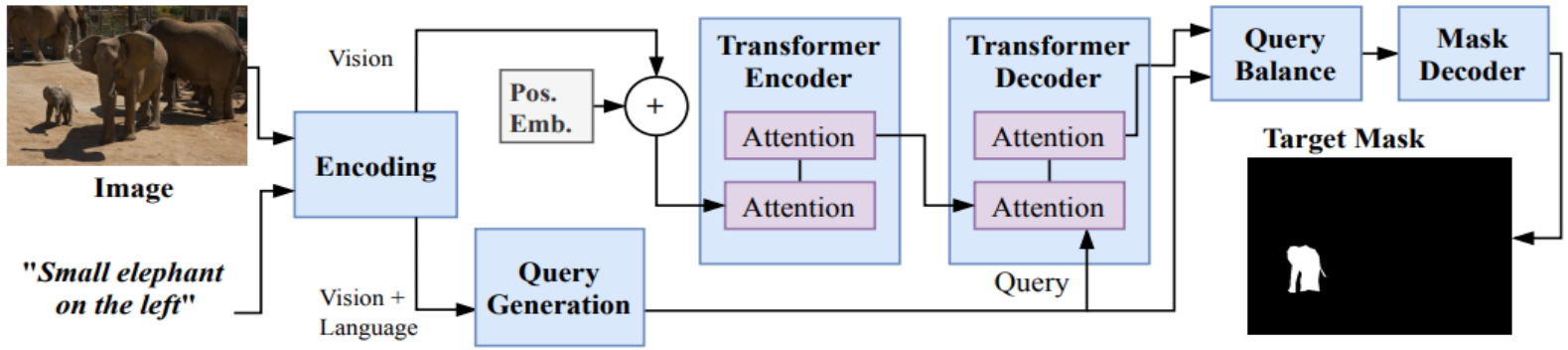
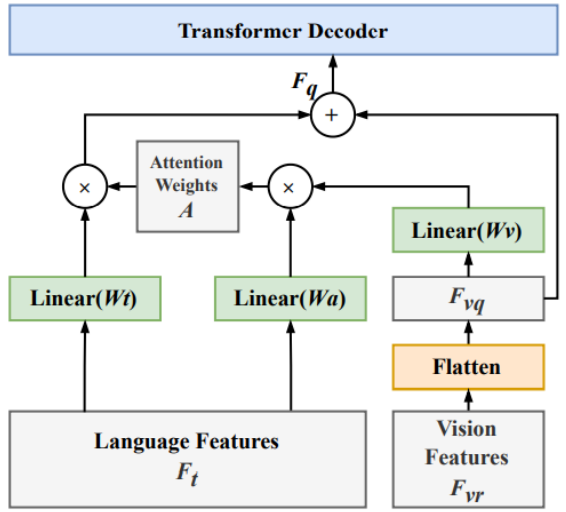
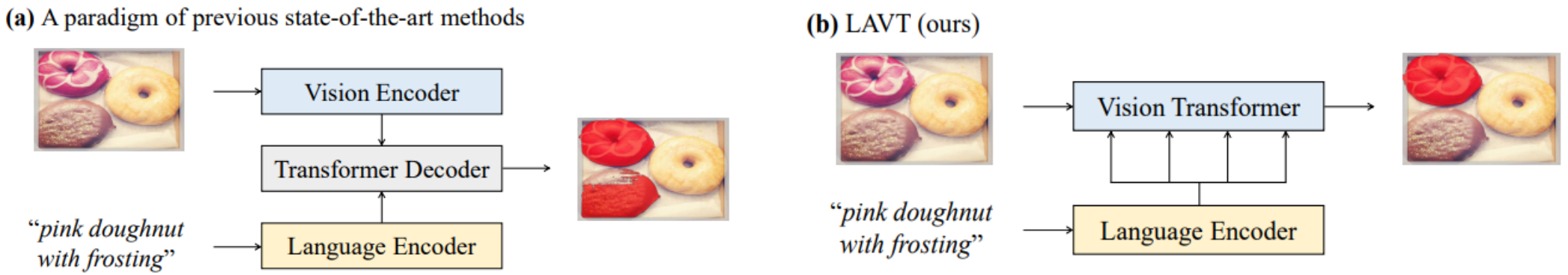


Figure 2. The overview of the network framework. Firstly, the input image and language expression are transformed into feature spaces. Features then processed by a transformer encoder-decoder model, generating a set of query responses. These responses are then decoded to output the target mask. "Pos. Emb.": Positional Embeddings.



Motivation

- In previous methods, cross-modal interactions occur only after feature encoding.
- The cross-modal decoder is solely responsible for aligning the visual and linguistic features.



- As a result, previous methods fail to effectively leverage the rich Transformer layers in the encoder for excavating helpful multi-modal context
- To address these issues, a potential solution is to exploit a visual encoder network for jointly embedding linguistic and visual features during visual encoding.

Swin Transformer

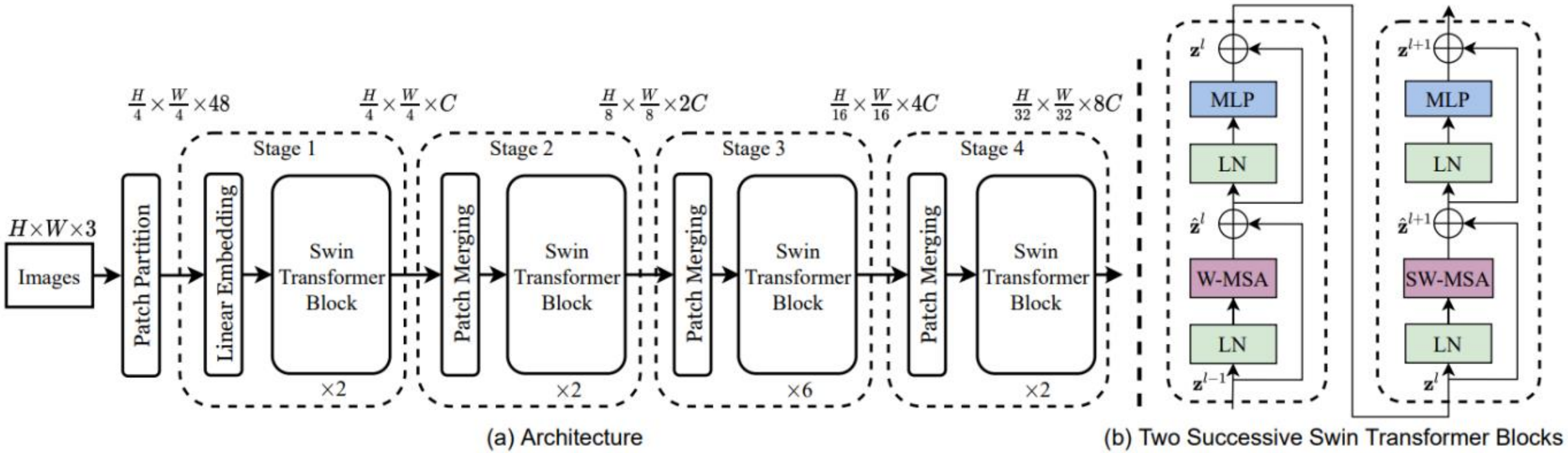


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

Contribution

- We propose LAVT (**L**anguage-**A**ware **V**ision **T**ransformer), a Transformer-based referring image segmentation framework that performs language-aware visual encoding in place of cross-modal fusion in a post-feature extraction step.
- We achieve new state-of-the-art results on three datasets for referring image segmentation, demonstrating the effectiveness and generality of the proposed method.

EFN [14]	62.76	65.69	59.67	51.50	55.24	43.01	-	-	51.93
BUSNet [55]	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56
CGAN [37]	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
LTS [25]	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
VLT [12]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76
LAVT (Ours)	**72.73**	**75.82**	**68.79**	**62.14**	**68.38**	**55.10**	**61.24**	**62.09**	**60.50**

Approach

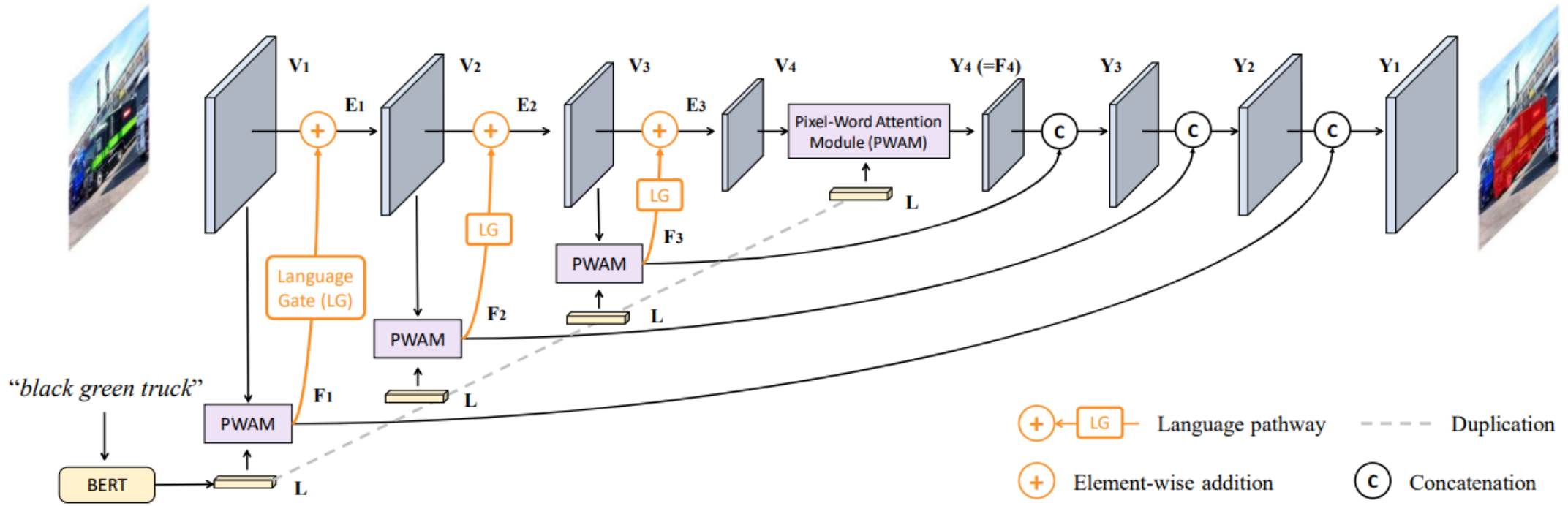
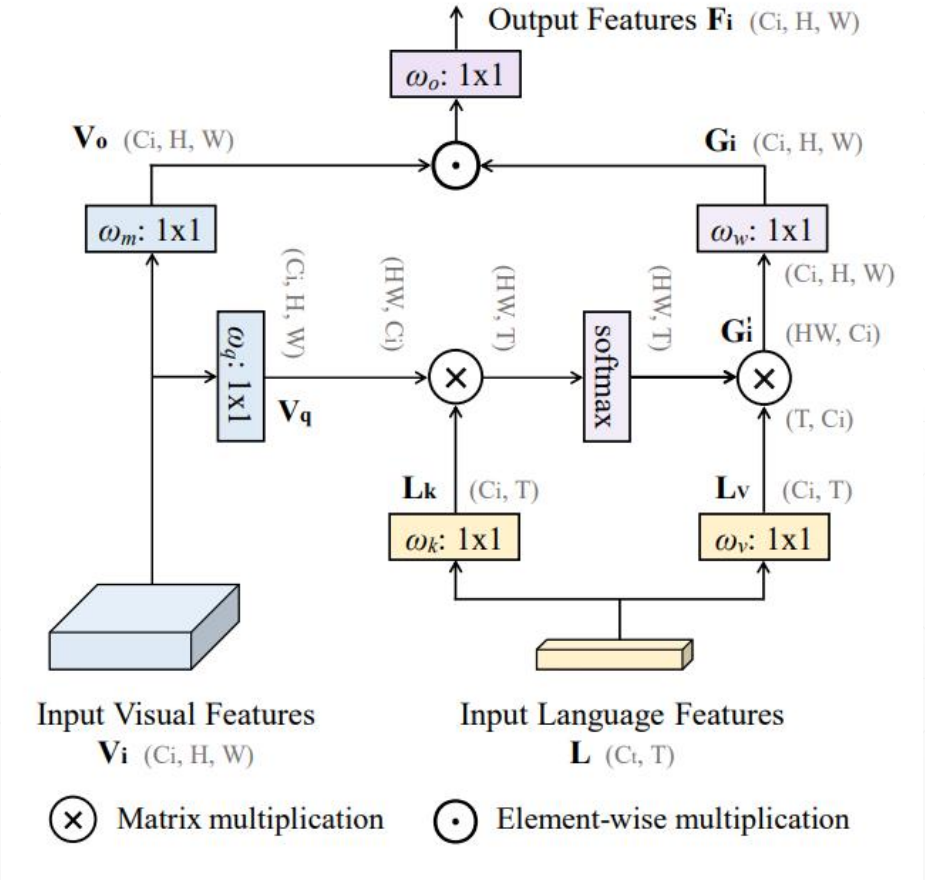


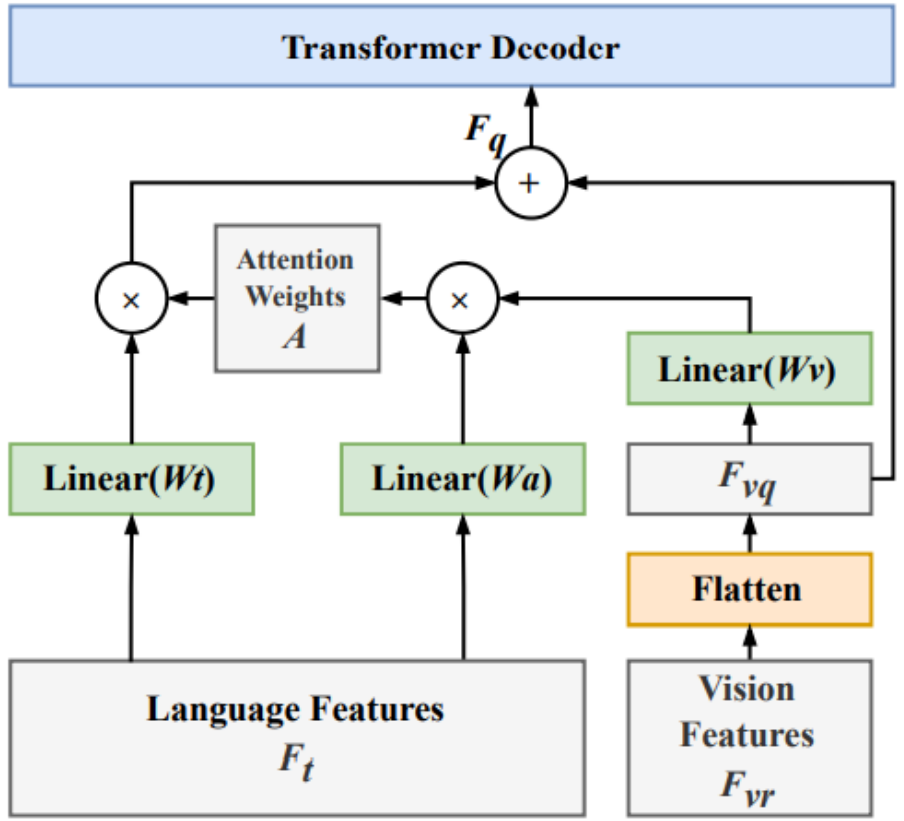
Figure 2. Overall pipeline of the proposed LAVT. We leverage a hierarchical vision Transformer [32] to perform language-aware visual encoding. At each stage, visual feature maps V_i, $i \in \{1, 2, 3, 4\}$ are encoded from the corresponding stage of Transformer layers (which are described in Sec. 3.1 and for diagrammatic clarity, are not illustrated in this figure). Then V_i are used as queries for generating a set of position-specific language feature maps F_i, $i \in \{1, 2, 3, 4\}$ in the pixel-word attention module (Sec. 3.2). Next, we adaptively fuse F_i with the original V_i via a language pathway (Sec. 3.3). The new visual feature maps E_i, $i \in \{1, 2, 3\}$ are then passed into the next stage of Transformer layers for further processing. A standard segmentation decoder head (Sec. 3.4) produces the final segmentation output.

Approach

- Pixel-Word Attention Module (LAVT)



- Query Generation Module (VLT)



Approach

- Language Pathway

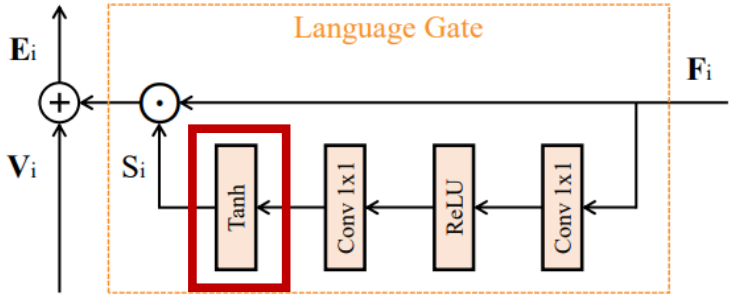
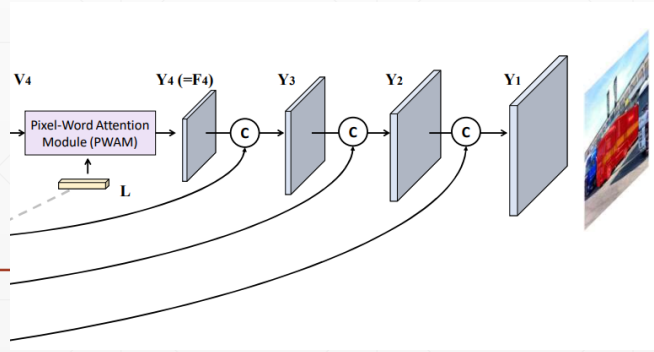
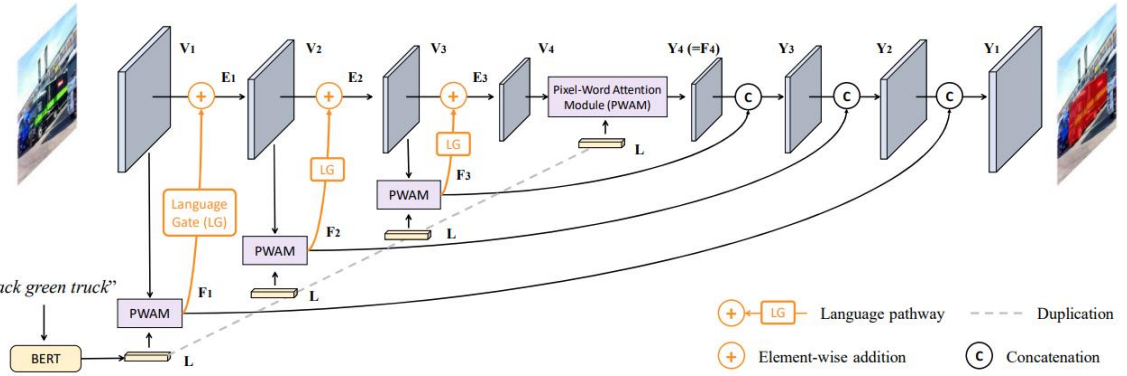


Figure 4. The schema of the language pathway, which leverages a language gate (LG) for controlling multi-modal information flow. LG is implemented as a two-layer perceptron.

- Segmentation Decoder-side





We combine the multi-modal feature maps, F_i, $i \in \{1, 2, 3, 4\}$, in a top-down manner to exploit multi-scale semantics for final segmentation. The decoding process can be described by the following recursive function

$$\begin{cases} Y_4 &= F_4, \\ Y_i &= \rho_i([\upsilon(Y_{i+1}); F_i]), \quad i = 3,\ 2,\ 1. \end{cases} \tag{10}$$

Here '[;]' denotes feature concatenation along the channel dimension, υ represents upsampling via bilinear interpolation, and ρ_i is a projection function implemented as two 3×3 convolutions connected by batch normalization [24] and ReLU [42] nonlinearity. The final feature maps, Y_1, are projected into two class score maps via a 1×1 convolution.

Experiments

	RefCOCO			RefCOCO+			G-Ref		
	val	test A	test B	val	test A	test B	val (U)	test (U)	val(G)
DMN [40]	49.78	54.83	45.13	38.88	44.22	32.29	-	-	36.76
RRN [28]	55.33	57.26	53.93	39.75	42.15	36.11	-	-	36.45
MAttNet [60]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-
CMSA [59]	58.32	60.61	55.09	43.76	47.60	37.89	-	-	39.98
CAC [7]	58.90	61.77	53.81	-	-	-	46.37	46.95	44.32
STEP [4]	60.04	63.46	57.97	48.19	52.33	40.41	-	-	46.40
BRINet [20]	60.98	62.99	59.21	48.17	52.32	42.11	-	-	48.04
CMPC [21]	61.36	64.53	59.64	49.56	53.44	43.23	-	-	49.05
LSCM [23]	61.47	64.99	59.55	49.34	53.12	43.50	-	-	48.05
CMPC+ [31]	62.47	65.08	60.82	50.25	54.04	43.47	-	-	49.89
MCN [38]	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
EFN [14]	62.76	65.69	59.67	51.50	55.24	43.01	-	-	51.93
BUSNet [55]	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56
CGAN [37]	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
LTS [25]	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
VLT [12]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76
LAVT (Ours)	**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 on three standard benchmark datasets. U: The UMD partition. G: The Google partition.

Experiments--Validity of LP and PWAM

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

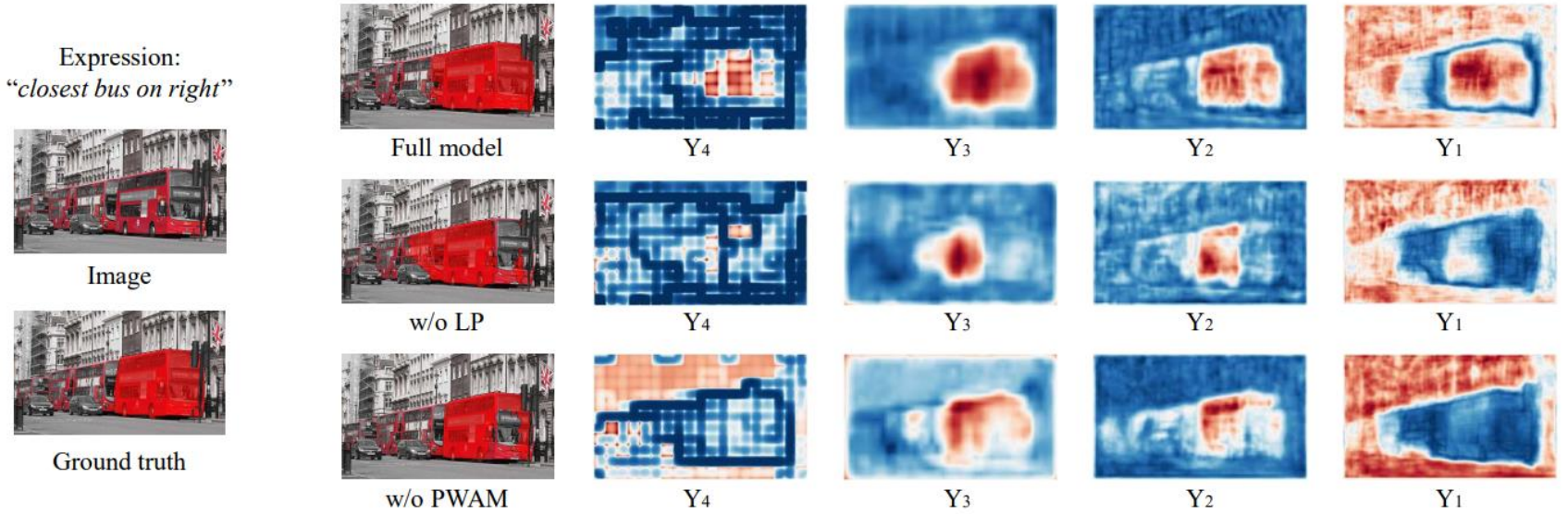
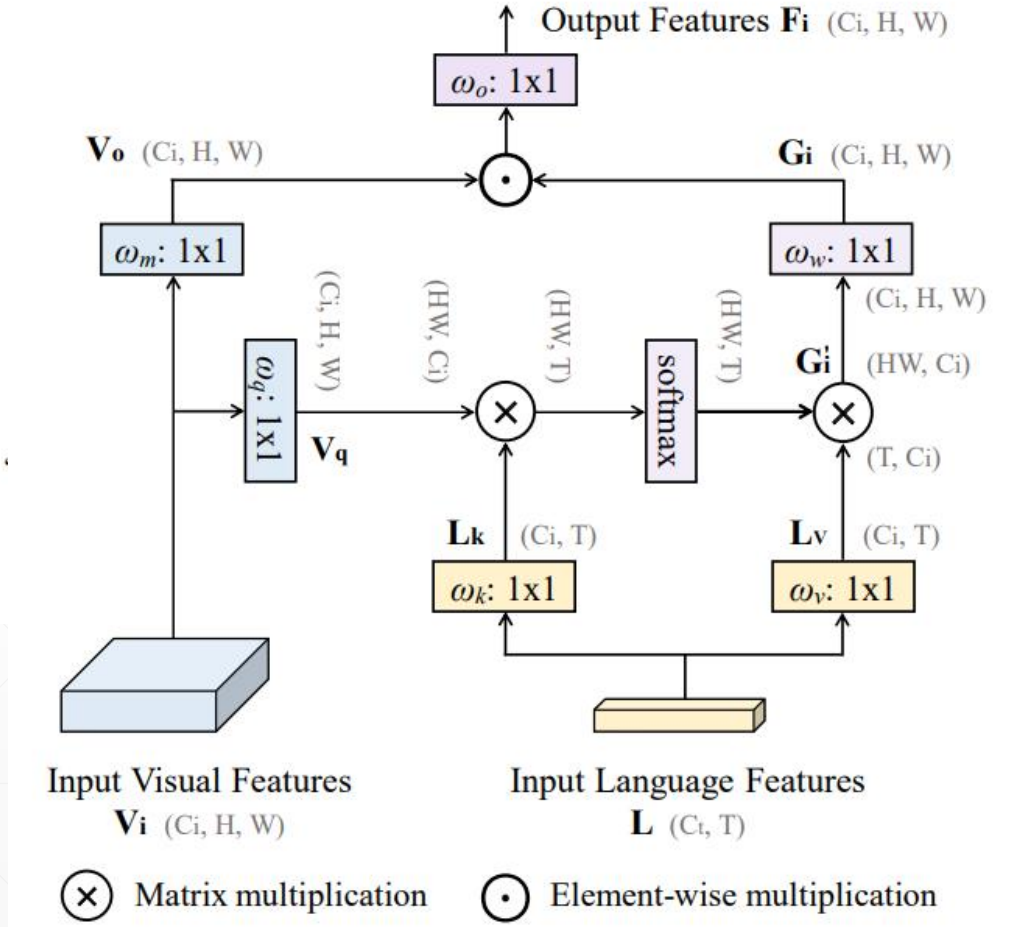
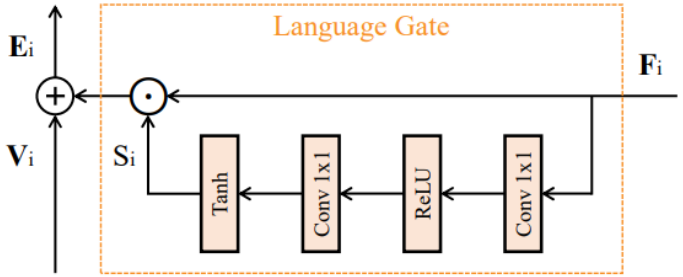


Figure 5. Visualized predictions and feature maps on an example from the RefCOCO validation set. From top to bottom, the left-most column illustrates the input expression, the input image, and the ground-truth mask overlaid on the input image. In each row, we visualize the predicted mask and the feature maps used for final classification (*i.e.*, Y_4, Y_3, Y_2, and Y_1) from left to right. LP represents the language pathway and PWAM represents the pixel-word attention module.

Experiments—Other Ablations





	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

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.

Experiments—Comparisons with VLT and EFN on a Fair Ground

Method	P@0.5	P@0.7	P@0.9	oIoU	mIoU
EFN (Swin-B)† [14]	82.55	73.27	31.68	70.76	72.95
VLT (Swin-B) [12]	83.24	72.81	24.64	70.89	71.98
Ours + VLT [12]	**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, VLT [12], and EFN [14] under exactly the same settings and using Swin-B [32] as the visual backbone on the RefCOCO validation set.

Method	RefCOCO@val
EFN	62.76
EFN+Swin	70.76
VLT	65.65
VLT+Swin	70.89
LAVT+VLT(Decoder)	72.12
LAVT	72.73

Visualization

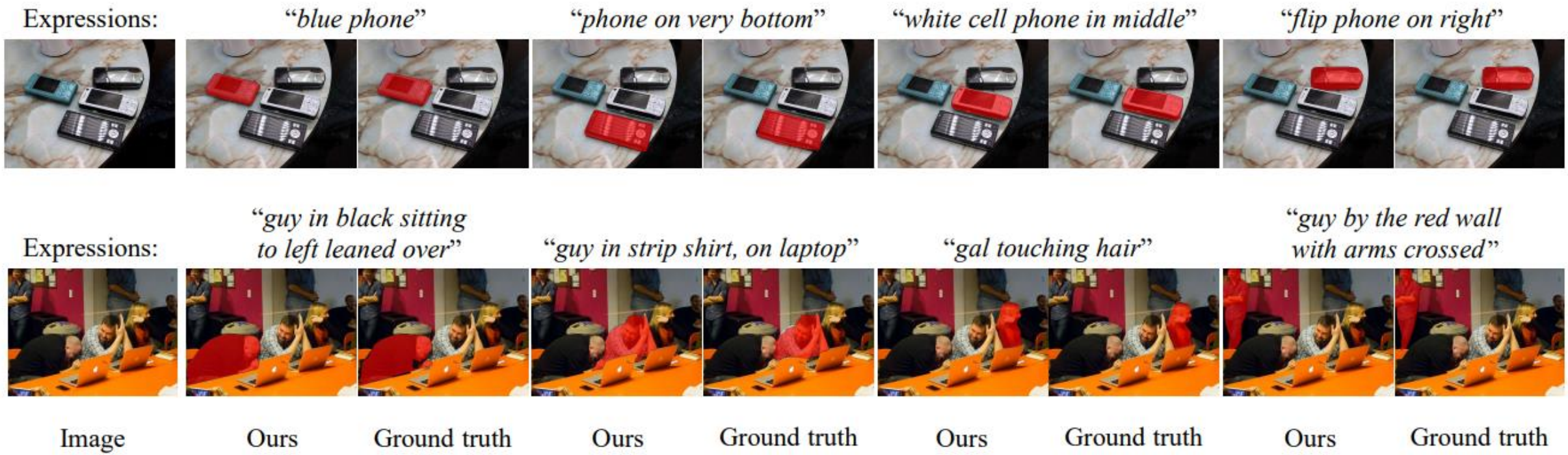


Figure 6. Visualizations of our predicted masks and the ground-truth masks on two examples from the RefCOCO validation set.

Rethinking

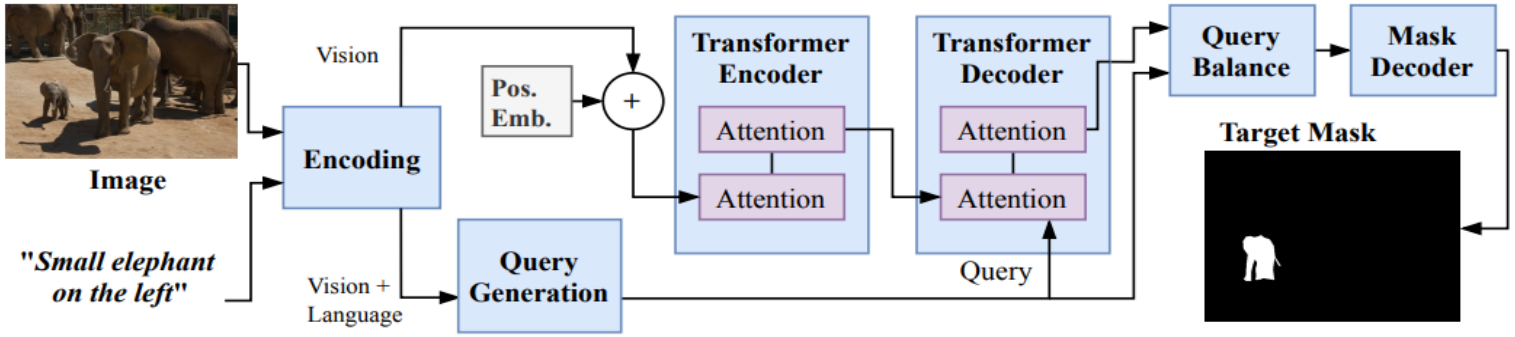
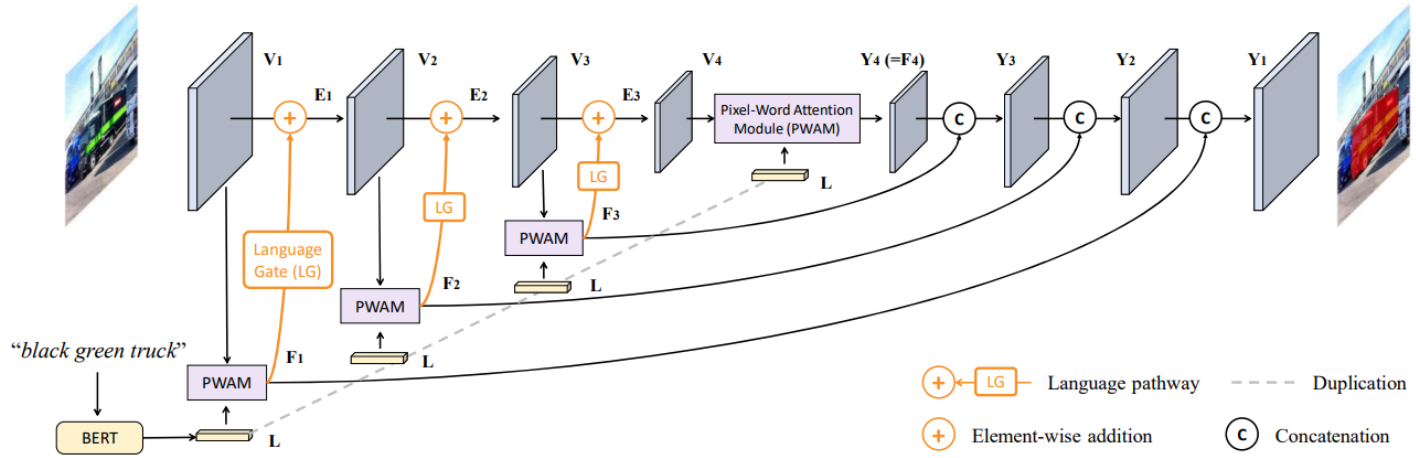
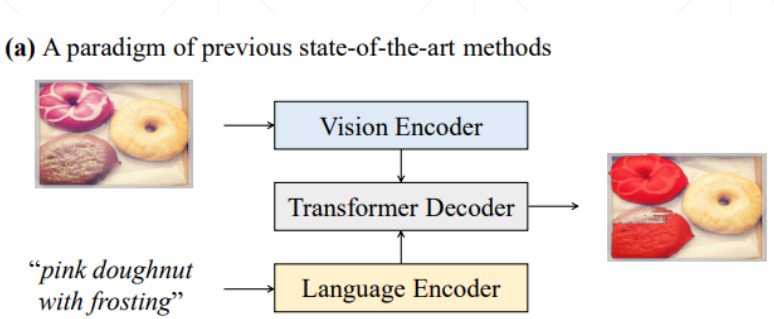
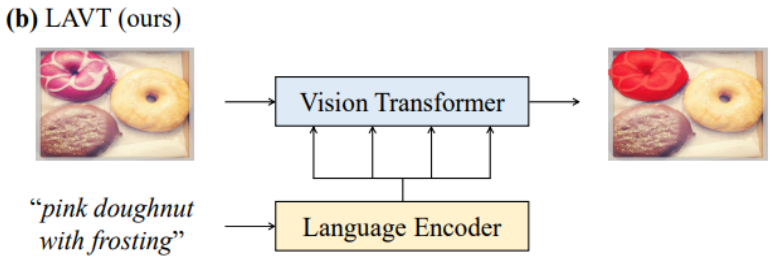


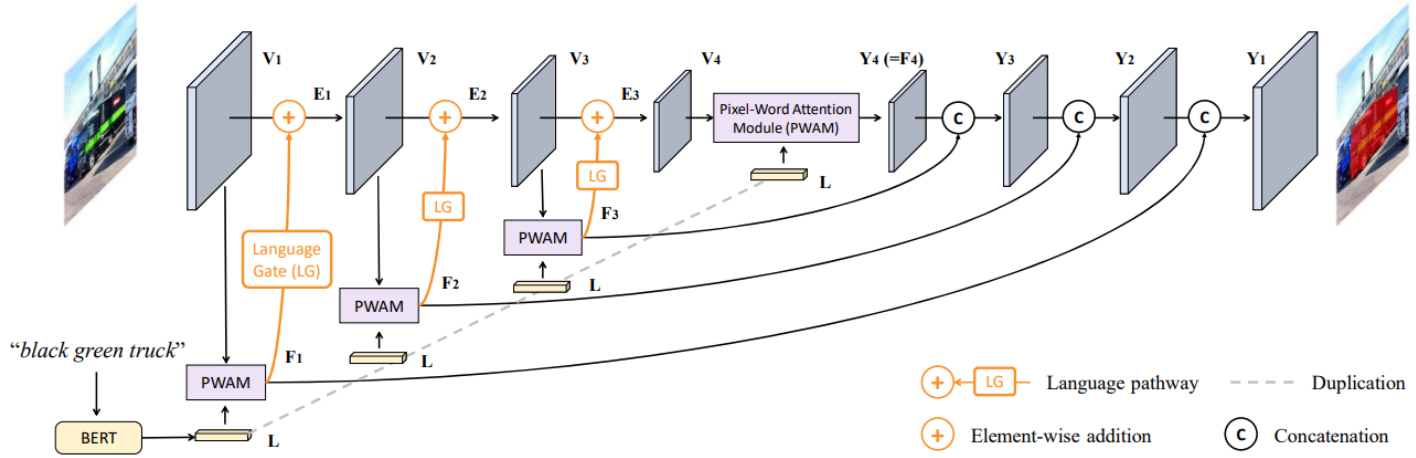
Image
"Small elephant on the left"
Vision
Encoding
Vision + Language
Pos. Emb.
Transformer Encoder
Attention
Attention
Transformer Decoder
Attention
Attention
Query
Query Generation
Query Balance
Mask Decoder
Target Mask

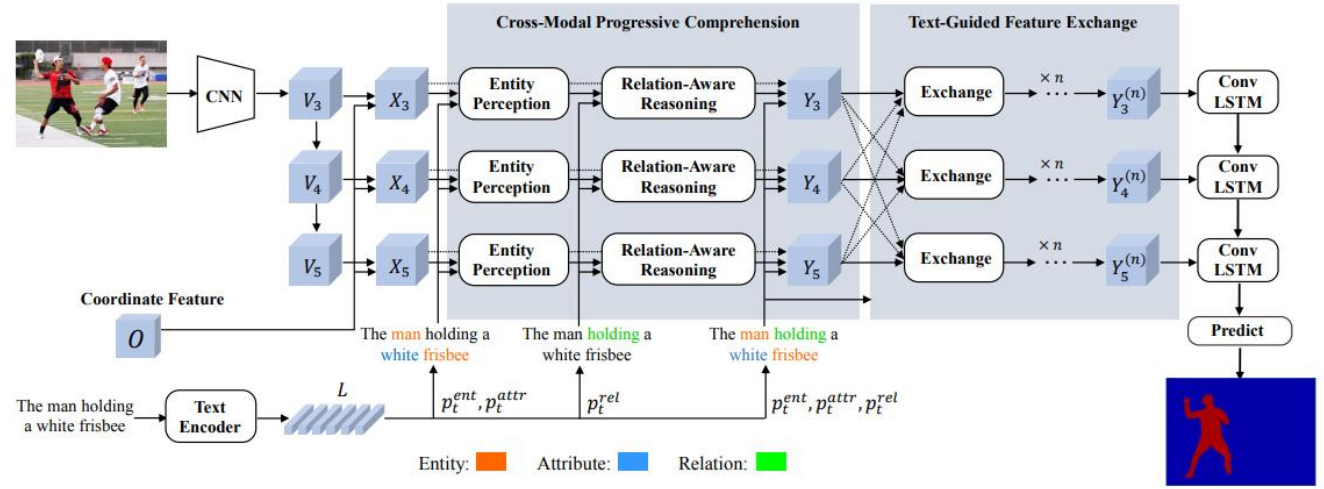

"black green truck"
BERT
PWAM
Language Gate (LG)
LG
PWAM
LG
PWAM
Pixel-Word Attention Module (PWAM)
PWAM
V1
V2
V3
V4
Y4 (=F4)
Y3
Y2
Y1
E1
E2
E3
F1
F2
F3
L
Language pathway
Duplication
Element-wise addition
Concatenation


(a) A paradigm of previous state-of-the-art methods
"pink doughnut with frosting"
Vision Encoder
Transformer Decoder
Language Encoder


(b) LAVT (ours)
"pink doughnut with frosting"
Vision Transformer
Language Encoder

Rethinking


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


Cross-Modal Progressive Comprehension
Text-Guided Feature Exchange
CNN
Entity Perception
Relation-Aware Reasoning
Exchange
Conv LSTM
Predict
Coordinate Feature
The man holding a white frisbee
Text Encoder
Entity:
Attribute:
Relation:

Thanks