CVPR 2022

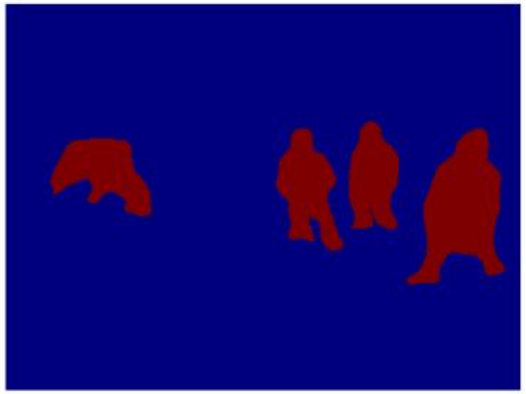
Referring-related Tasks

Mengxue

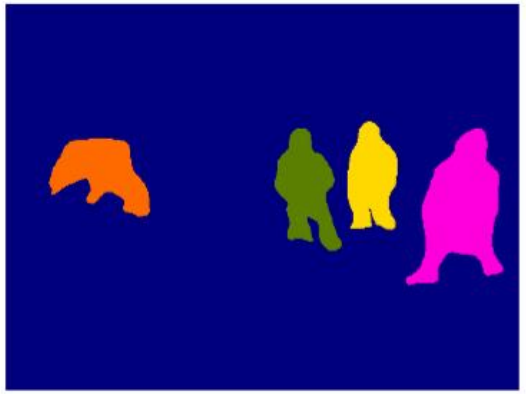
Introduction

- What is Referring? → 指代 → 用文本描述指代特定目标
- Referring Image Segmentation: given a sentence, segment corresponding instance.

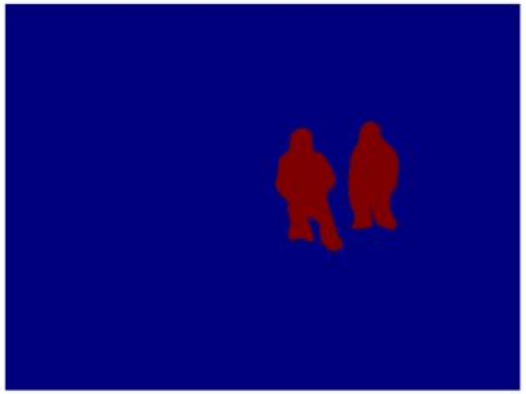
(a) input image



(b) object class segmentation of class ***people***



(c) object instance segmentation of class ***people***



(d) segmentation from expression *"people in blue coat"*

Introduction

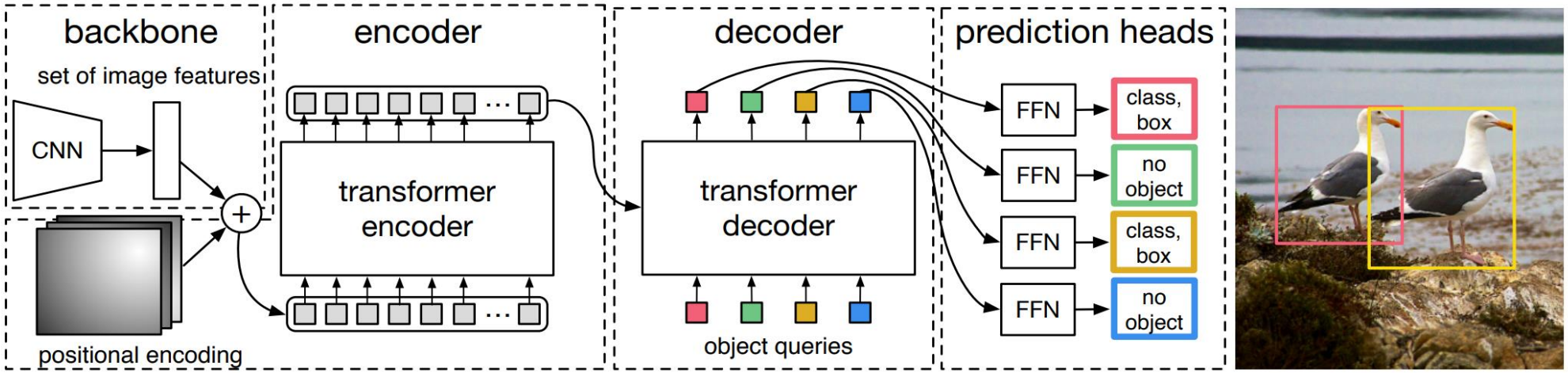
Referring Image Grounding	Detection
Referring Image Segmentation	Segmentation
Referring Video Segmentation	Video Segmentation
Unsupervised Referring Grounding	Unsupervised
Referring Image Matting	Matting
Language-driven Other Visual Tasks	Image Editing/Generation…

1/ DETR-like frameworks

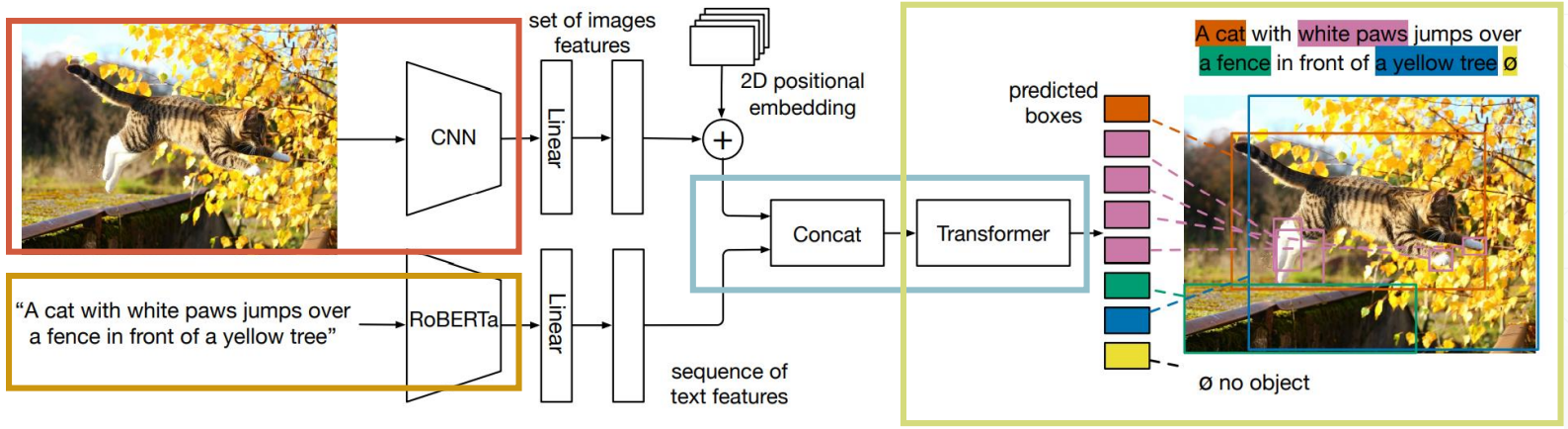
- **Referring Image Grounding**
  - Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning
- **Referring Image Segmentation**
  - CRIS: CLIP-driven Referring Image Segmentation
- **Referring Video Segmentation**
  - Language as Queries for Referring Video Object Segmentation

DETR-like frameworks

- DETR



- Vision Encoder
- Language Encoder
- V-L Interaction
- Detection Regression



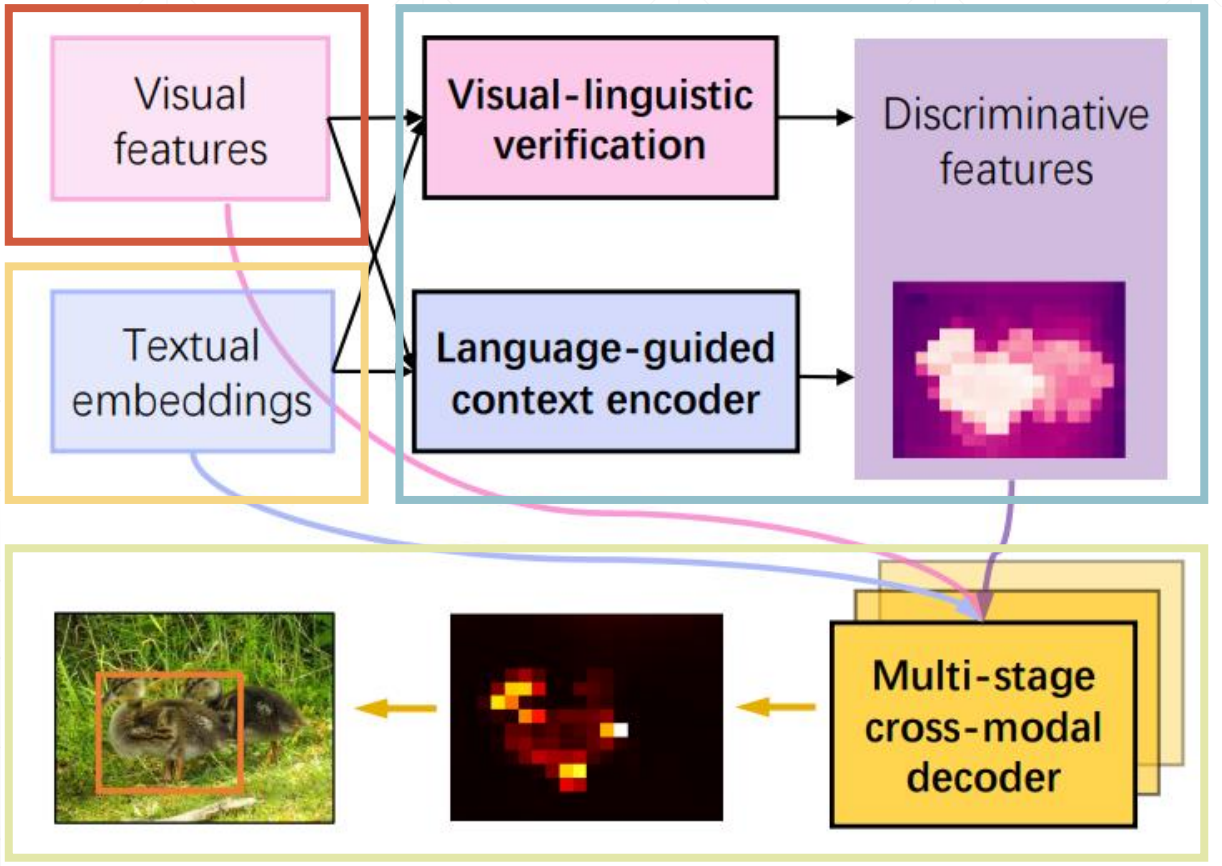
MDETR - Modulated Detection for End-to-End Multi-Modal Understanding, ICCV21

Referring Image Grounding

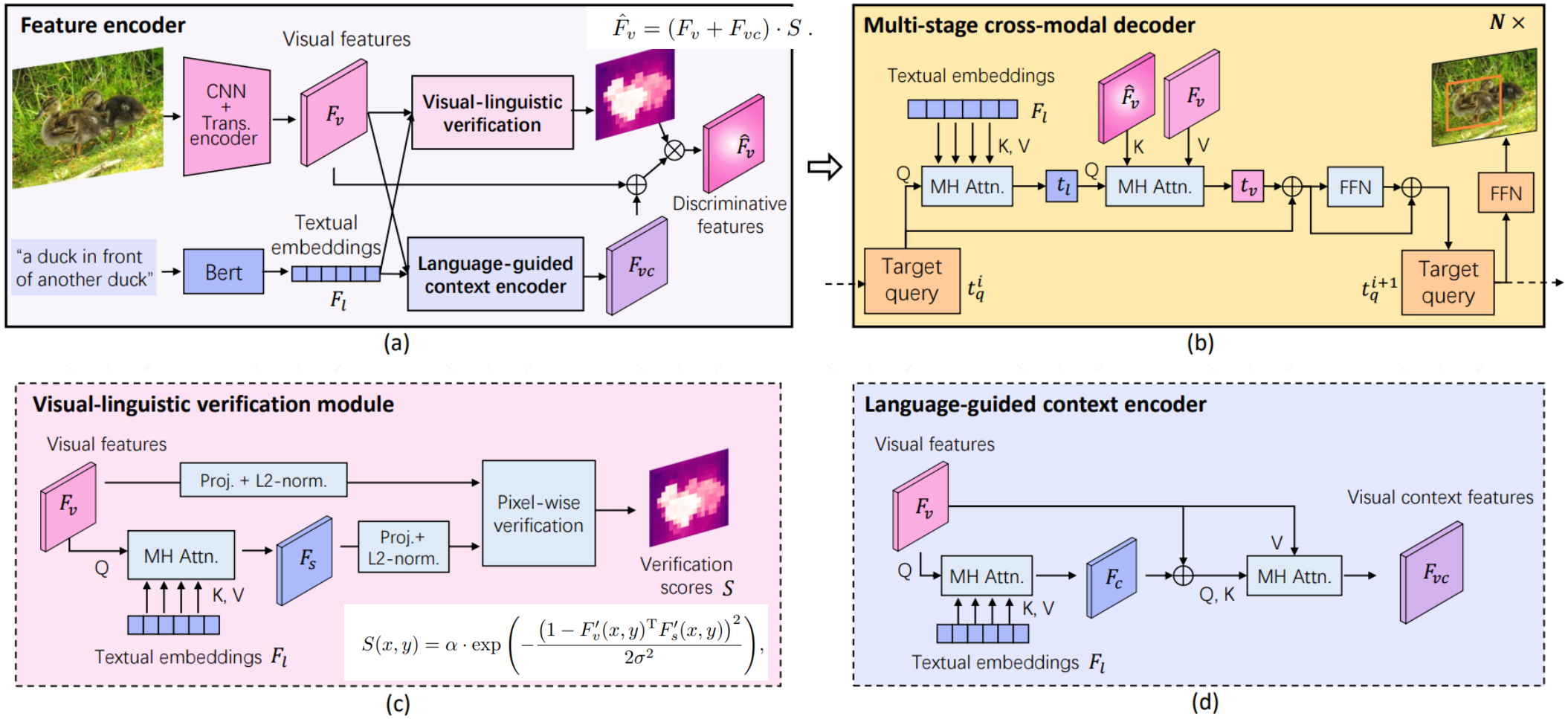
Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning

Li Yang[1,2*], Yan Xu[3*], Chunfeng Yuan[1†], Wei Liu[1], Bing Li[1], and Weiming Hu[1,2,4]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]The Chinese University of Hong Kong
[4]CAS Center for Excellence in Brain Science and Intelligence Technology
{li.yang,cfyuan,bli,wmhu}@nlpr.ia.ac.cn, yanxu@link.cuhk.edu.hk, liuwei@ia.ac.cn

- Vision Encoder
- Language Encoder
- V-L Interaction
- Detection Regression



Referring Image Grounding



Referring Image Grounding

Models	Backbone	RefCOCO			RefCOCO+			RefCOCOg		
		val	*testA*	*testB*	*val*	*testA*	*testB*	*val-g*	*val-u*	*test-u*
Two-stage:										
CMN [13]	VGG16	-	71.03	65.77	-	54.32	47.76	57.47	-	-
VC [47]	VGG16	-	73.33	67.44	-	58.40	5.318	62.30	-	-
ParalAttn [50]	VGG16	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MAttNet [44]	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
LGRANs [34]	VGG16	-	76.60	66.40	-	64.00	53.40	61.78	-	-
DGA [40]	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28
RvG-Tree [12]	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree [20]	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
Ref-NMS [2]	ResNet-101	80.70	84.00	76.04	68.25	73.68	59.42	-	70.55	70.62
One-stage:										
SSG [3]	DarkNet-53	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-
FAOA [43]	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [18]	DLA-34	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC-Large [42]	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
LBYL-Net [15]	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-
Transformer-based:										
TransVG [5]	ResNet-50	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44
TransVG [5]	ResNet-101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
VLTVG (ours)	ResNet-50	**84.53**	**87.69**	**79.22**	**73.60**	**78.37**	**64.53**	**72.53**	**74.90**	**73.88**
VLTVG (ours)	ResNet-101	**84.77**	**87.24**	**80.49**	**74.19**	**78.93**	**65.17**	**72.98**	**76.04**	**74.18**

Multi-stage Decooder	Context Encoder	V-L Verification	#params	GFLOPS	Acc (%)
			143.37M	41.10	63.64
✓			151.26M	41.39	66.02
✓	✓		151.79M	41.67	68.44
✓	✓	✓	152.18M	41.79	**71.62**

Table 4. Comparison of different decoder stages used to perform cross-modal reasoning for visual grounding.

The decoder stages (N)	#params	GFLOPS	Acc (%)
$N = 1$	143.37M	41.10	63.64
$N = 2$	144.95M	41.15	65.05
$N = 4$	148.10M	41.27	65.70
$N = 6$	151.26M	41.39	**66.02**
$N = 8$	154.42M	41.51	65.97

Table 5. Comparison of our method with the transformer-based approach for visual-linguistic feature learning.

The V-L feature learning	#params	GFLOPS	Acc (%)
None	151.26M	41.39	66.02
Trans. encoder layers ($\times 1$)	152.69M	42.08	69.37
Trans. encoder layers ($\times 2$)	154.00M	42.77	69.22
Trans. encoder layers ($\times 3$)	155.32M	43.46	69.15
Trans. encoder layers ($\times 4$)	156.63M	44.15	69.55
Ours (V-L verification + context)	152.18M	41.79	**71.62**

Referring Image Segmentation

CRIS: CLIP-Driven Referring Image Segmentation

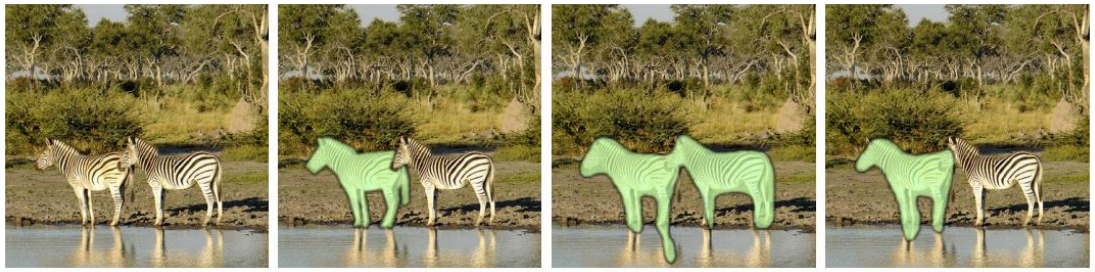
Zhaoqing Wang[1,2*] Yu Lu[3*] Qiang Li[4*] Xunqiang Tao[3] Yandong Guo[3]
Mingming Gong[5] Tongliang Liu[1]

[1]University of Sydney; [2]OPPO Research Institute; [3]Beijing University of Posts and Telecommunications
[4]Kuaishou Technology; [5]University of Melbourne

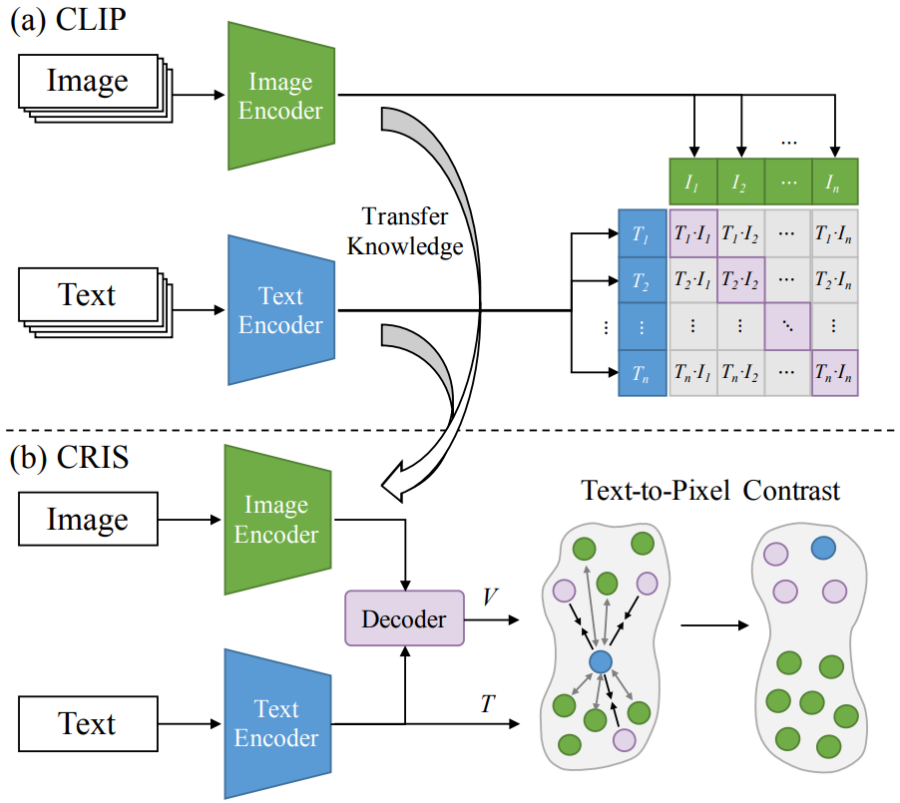
Language: *"a blond haired , blue eyed young boy in a blue jacket"*

(a) Image (b) GT (c) Naïve (f) Ours

Language: *"a zebra ahead of the other zebra"*



(a) Image (b) GT (c) Naïve (f) Ours

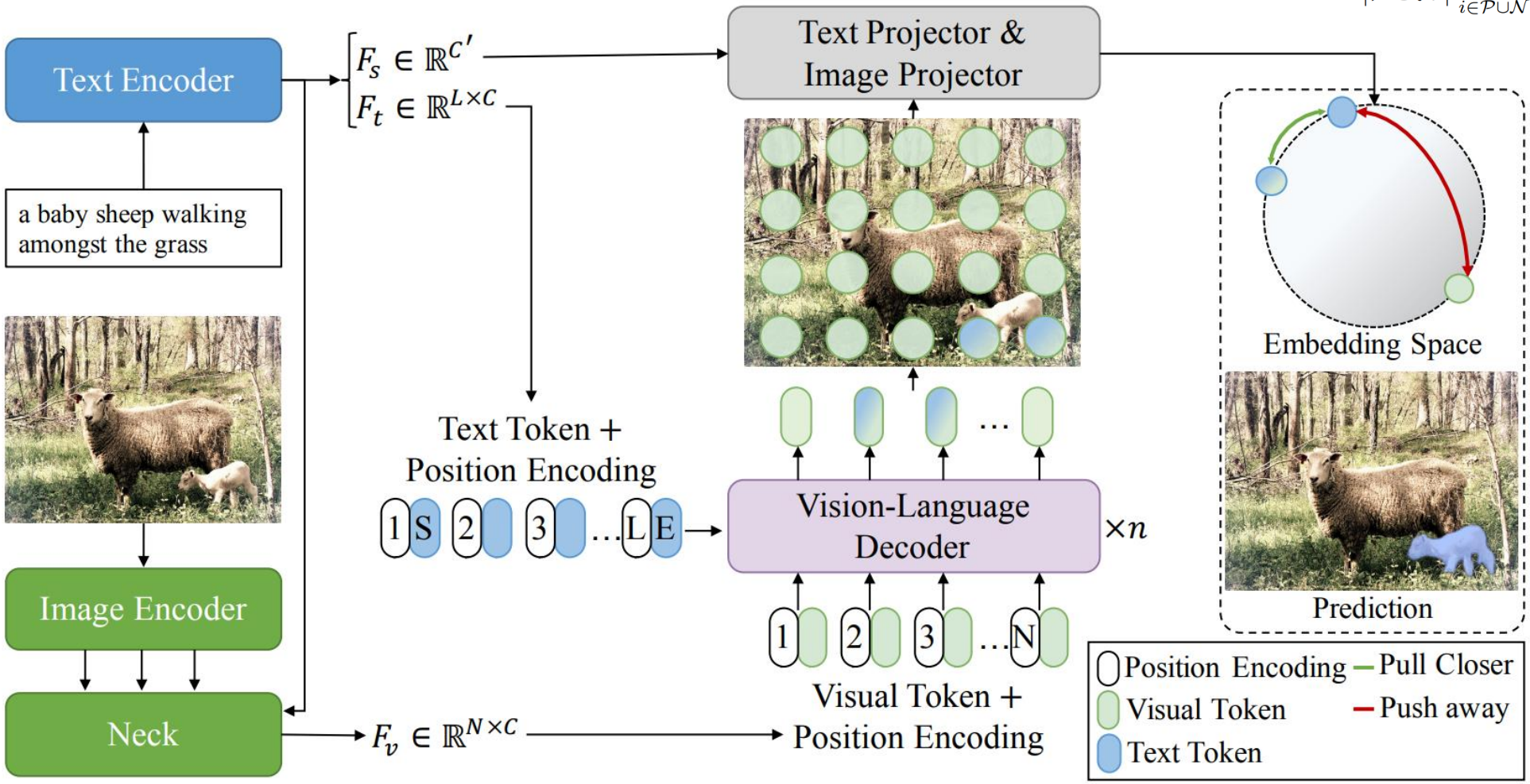


- Naïve: directly finetune CLIP
- {gt_mask, Sigmoid(pixel_feat·text_feat)}

Referring Image Segmentation

$$L^i_{con}(z_t, z^i_v) = \begin{cases} -\log \sigma(z_t \cdot z^i_v), & i \in \mathcal{P}, \\ -\log(1 - \sigma(z_t \cdot z^i_v)), & i \in \mathcal{N}, \end{cases} \quad (9)$$

$$L_{con}(z_t, z_v) = \frac{1}{|\mathcal{P} \cup \mathcal{N}|} \sum_{i \in \mathcal{P} \cup \mathcal{N}} L^i_{con}(z_t, z^i_v), \quad (10)$$



Referring Image Segmentation

Method	Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	test A	test B	val	test A	test B	val	test
RMI* [25]	ResNet-101	45.18	45.69	45.57	29.86	30.48	29.50	-	-
DMN [33]	ResNet-101	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN* [22]	ResNet-101	55.33	57.26	53.95	39.75	42.15	36.11	-	-
MAttNet [50]	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [26]	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA* [49]	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2Seg [5]	ResNet-101	58.90	61.77	53.81	-	-	-	46.37	46.95
BCAN* [16]	ResNet-101	61.35	63.37	59.57	48.57	52.87	42.13	-	-
CMPC* [17]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-
LSCM* [18]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	-	-
MCN [30]	DarkNet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CGAN [29]	DarkNet-53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
EFNet [8]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
LTS [19]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [6]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
CRIS (Ours)	ResNet-50	69.52	72.72	64.70	61.39	67.10	52.48	59.35	59.39
CRIS (Ours)	ResNet-101	**70.47**	**73.18**	**66.10**	**62.27**	**68.08**	**53.68**	**59.87**	**60.36**

Referring Image Segmentation

Dataset	*Con.*	*Dec.*	*n*	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90	Params	FPS
RefCOCO	-	-	-	62.66	72.55	67.29	59.53	43.52	12.72	131.86	27.30
	✓	-	-	64.64	74.89	69.58	61.70	45.50	13.31	134.22	25.79
	-	✓	1	66.31	77.66	72.99	65.67	48.43	14.81	136.07	23.02
	✓	✓	1	68.66	80.16	75.72	68.82	51.98	15.94	138.43	22.64
	✓	✓	2	69.13	80.96	76.60	69.67	52.23	16.09	142.64	20.68
	✓	✓	3	**69.52**	**81.35**	**77.54**	**70.79**	**52.65**	16.21	146.85	19.22
	✓	✓	4	69.18	80.99	76.74	69.32	52.57	**16.37**	151.06	18.26
RefCOCO+	-	-	-	50.17	54.55	47.69	40.19	28.75	8.21	131.86	27.30
	✓	-	-	53.15	58.28	53.74	46.67	34.01	9.30	134.22	25.79
	-	✓	1	54.73	63.31	58.89	52.46	38.53	11.70	136.07	23.02
	✓	✓	1	59.97	69.19	64.85	58.17	43.47	13.39	138.43	22.64
	✓	✓	2	60.75	70.69	66.83	60.74	45.69	13.42	142.64	20.68
	✓	✓	3	**61.39**	**71.46**	**67.82**	**61.80**	**47.00**	**15.02**	146.85	19.22
	✓	✓	4	61.15	71.05	66.94	61.25	46.98	14.97	151.06	18.26
G-Ref	-	-	-	49.24	53.33	45.49	36.58	23.90	6.92	131.86	25.72
	✓	-	-	52.67	59.27	52.45	44.12	29.53	8.80	134.22	25.33
	-	✓	1	51.46	58.68	53.33	45.61	31.78	10.23	136.07	22.57
	✓	✓	1	57.82	66.28	60.99	53.21	38.58	13.38	138.43	22.34
	✓	✓	2	58.40	67.30	61.72	54.70	39.67	13.40	142.64	20.61
	✓	✓	3	**59.35**	**68.93**	**63.66**	**55.45**	**40.67**	**14.40**	146.85	19.14
	✓	✓	4	58.79	67.91	63.11	55.43	39.81	13.48	151.06	17.84

Referring Image Segmentation

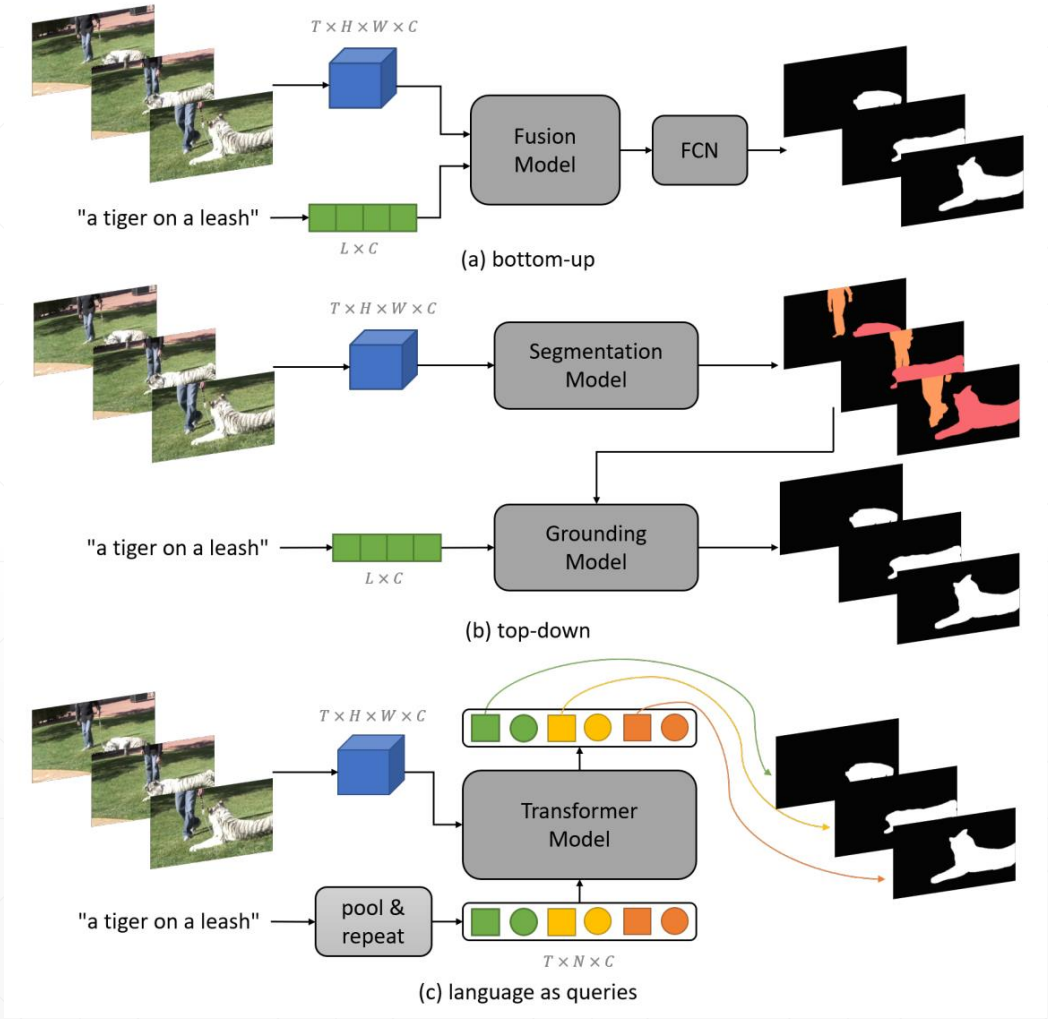
Figure 5. **Qualitative examples of failure cases.** *Best viewed in color.*

Referring Video Segmentation

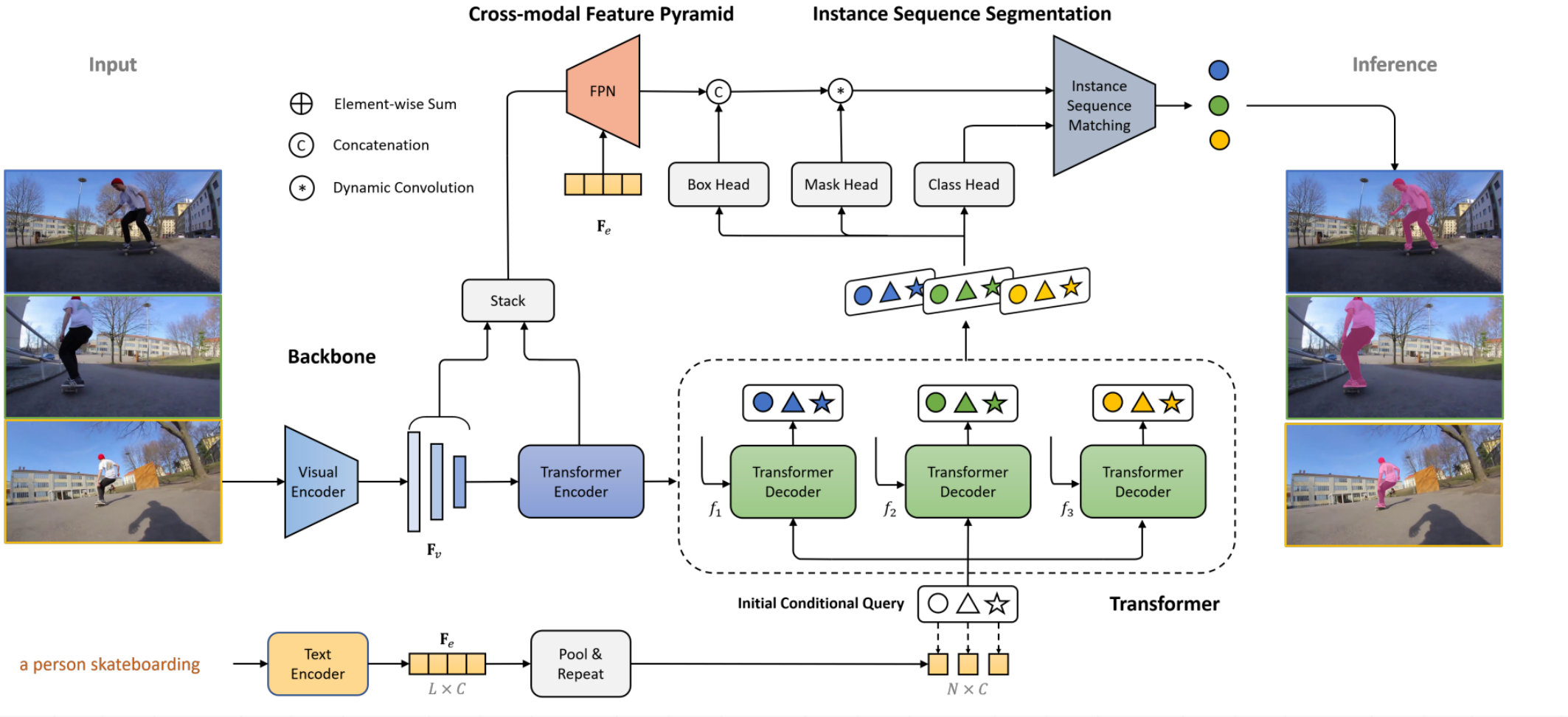
Language as Queries for Referring Video Object Segmentation

Jiannan Wu[1], Yi Jiang[2] Peize Sun[1], Zehuan Yuan[2], Ping Luo[1]
[1]The University of Hong Kong [2]ByteDance

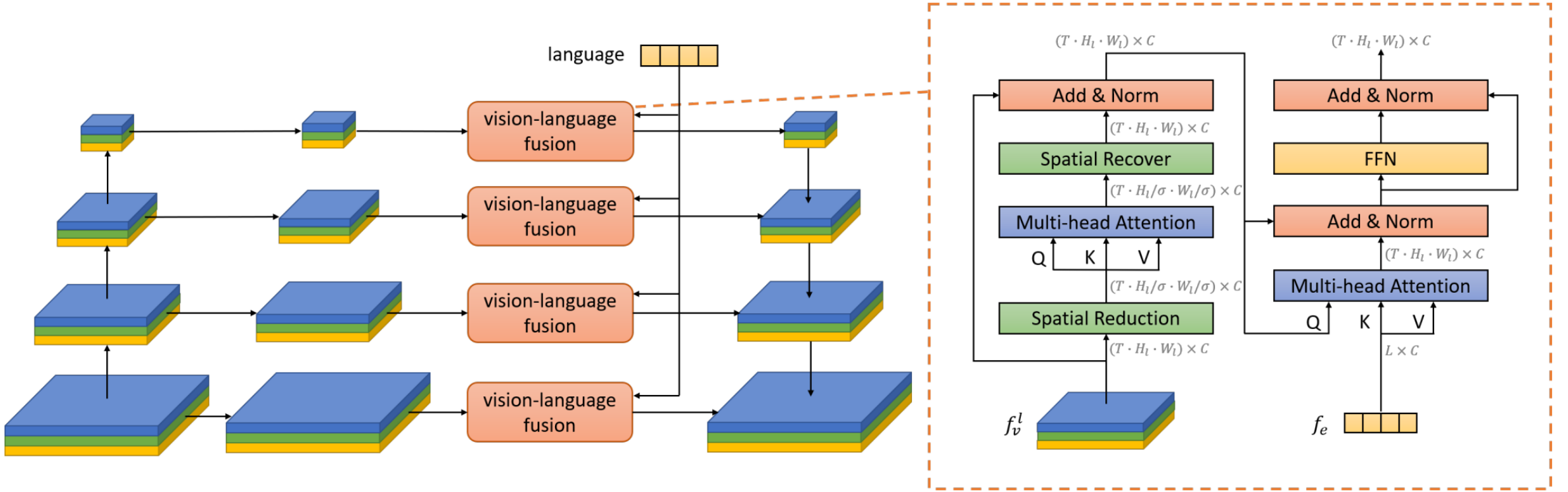
- Bottom-up
  - Early V-L merge + FCN-like segment
- Top-down
  - Instance segment + Late V-L merge
  - Pick out the instance mask
- Languages as queries
  - Early V-L merge in DETR-like model



Referring Video Segmentation



Referring Video Segmentation



Referring Video Segmentation

Method	Backbone	Ref-Youtube-VOS			Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Spatial Visual Backbones							
CMSA [49]	ResNet-50	34.9	33.3	36.5	34.7	32.2	37.2
CMSA + RNN [49]	ResNet-50	36.4	34.8	38.1	40.2	36.9	43.5
URVOS [37]	ResNet-50	47.2	45.3	49.2	51.5	47.3	56.0
ReferFormer	ResNet-50	55.6	54.8	56.5	**58.5**	**55.8**	**61.3**
ReferFormer + CFBI	ResNet-50	**59.4**	**58.1**	**60.8**	-	-	-
PMINet [7]	ResNeSt-101	48.2	46.7	49.6	-	-	-
PMINet* [7]	ResNeSt-101	53.0	51.5	54.5	-	-	-
CITD* [20]	ResNet-101	56.4	54.8	58.1	-	-	-
ReferFormer	ResNet-101	57.3	56.1	58.4	-	-	-
ReferFormer*	ResNet-101	**60.3**	**58.8**	**61.8**	-	-	-
PMINet* [7]	Ensemble	54.2	53.0	55.5	-	-	-
CITD* [20]	Ensemble	61.4	60.0	62.7	-	-	-
ReferFormer	Swin-L	62.4	60.8	64.0	**60.5**	**57.6**	**63.4**
ReferFormer*	Swin-L	**63.3**	**61.6**	**65.1**	-	-	-
Spatio-temporal Visual Backbones							
MTTR† ($\omega = 12$) [2]	Video-Swin-T	55.3	54.0	56.6	-	-	-
ReferFormer † ($\omega = 5$)	Video-Swin-T	**55.8**	**54.8**	**56.9**	-	-	-
ReferFormer	Video-Swin-T	59.4	58.0	60.9	-	-	-
ReferFormer	Video-Swin-S	60.1	58.6	61.6	-	-	-
ReferFormer	Video-Swin-B	**62.9**	**61.3**	**64.6**	**61.1**	**58.1**	**64.1**

Components	$\mathcal{J}$	$\mathcal{F}$
Baseline	47.2 (−7.6)	50.1 (−6.8)
w/o Visual-language Fusion	53.0 (−1.8)	56.2 (−0.7)
w/o Relative Coordinates	53.7 (−1.1)	55.9 (−1.0)
Full Model	54.8	56.9

Table 4. Ablation study on the components of ReferFormer.

Queries	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	53.6	52.7	54.5
3	54.2	53.2	55.2
5	55.8	54.8	56.9
8	55.3	54.1	56.6

(a) The effect of query number.

Frames	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	50.0	48.4	51.6
3	54.8	53.6	56.0
5	55.8	54.8	56.9

(b) The effect of frame number.

Class	Box	Mask	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✓	✓		55.2	54.0	56.4
✓		✓	54.5	53.5	55.5
✓	✓	✓	55.8	54.8	56.9

(c) The effect of label assignment method.

Region similarity (J) / Contour accuracy (F)

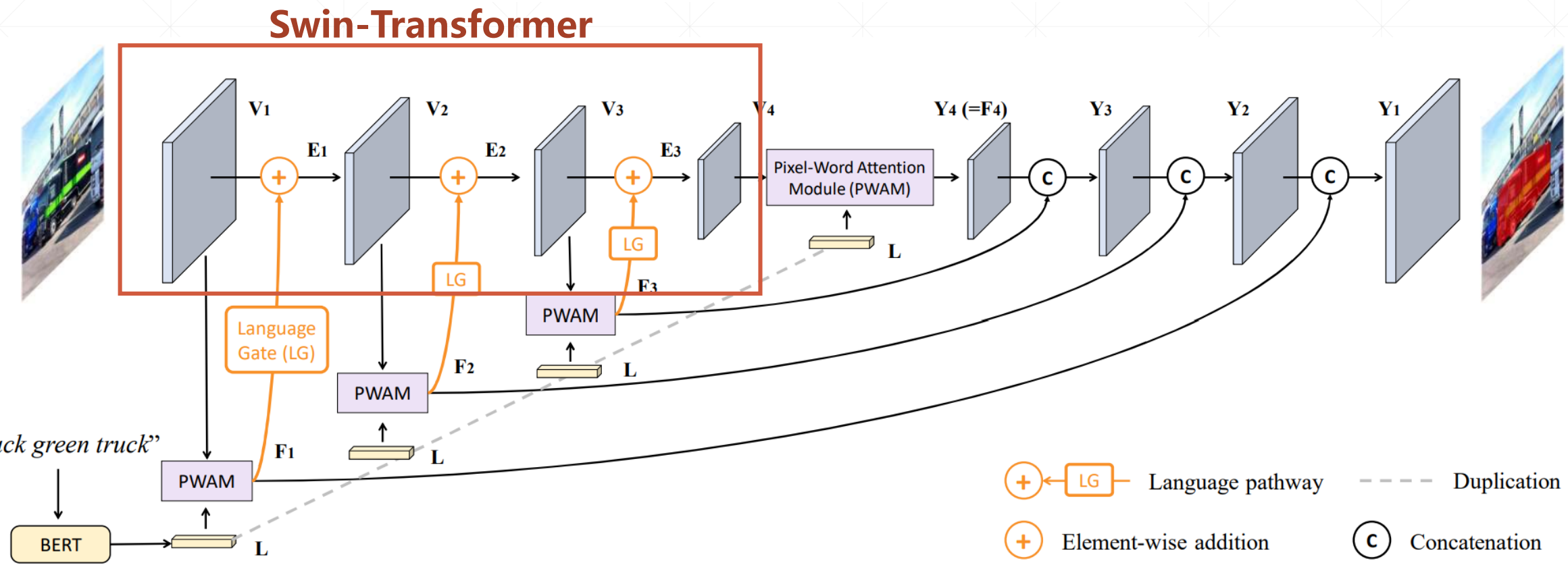
2/ Swin-Transformer-combined frameworks

- **Referring Image Segmentation**
- *(CVPR22)* LAVT: Image-Aware Vision Transformers for Referring Image Segmentation
- **Referring Image Grounding**
- *(ICCV21)* TransVG: End-to-End Visual Grounding with Transformers
- *(CVPR22)* Shifting More Attention to Visual Backbone: Query-modulated Refinement Network for End-to-End Visual Grounding

Referring Image Segmentation

LAVT: Language-Aware Vision Transformer for Referring Image Segmentation

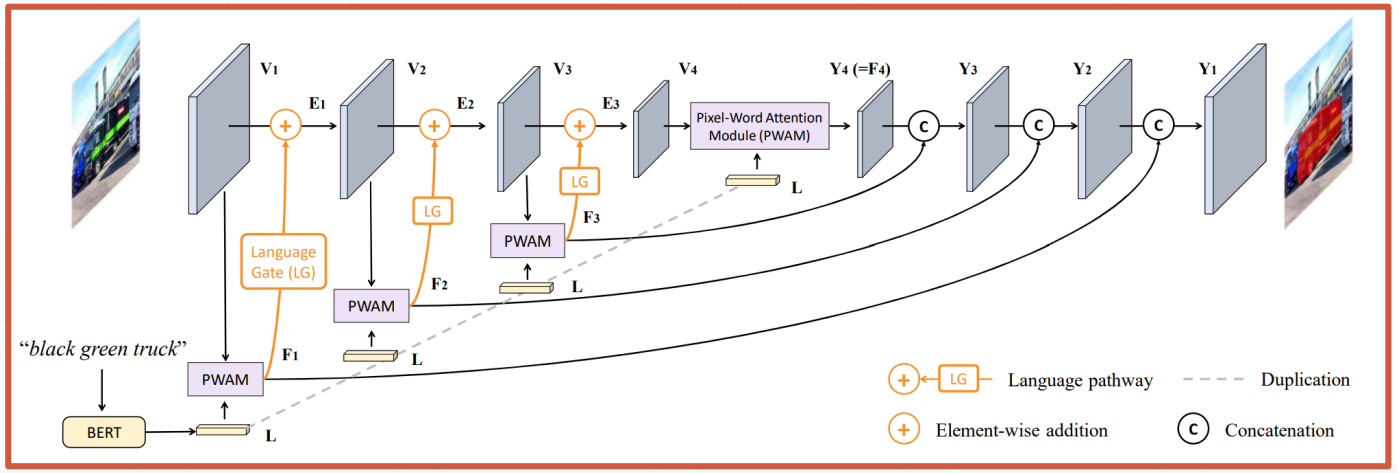
Zhao Yang[1], Jiaqi Wang[2], Yansong Tang[1], Kai Chen[2,3], Hengshuang Zhao[1,4], Philip H.S. Torr[1]
[1]University of Oxford, [2]Shanghai AI Laboratory,
[3]SenseTime Research, [4]The University of Hong Kong



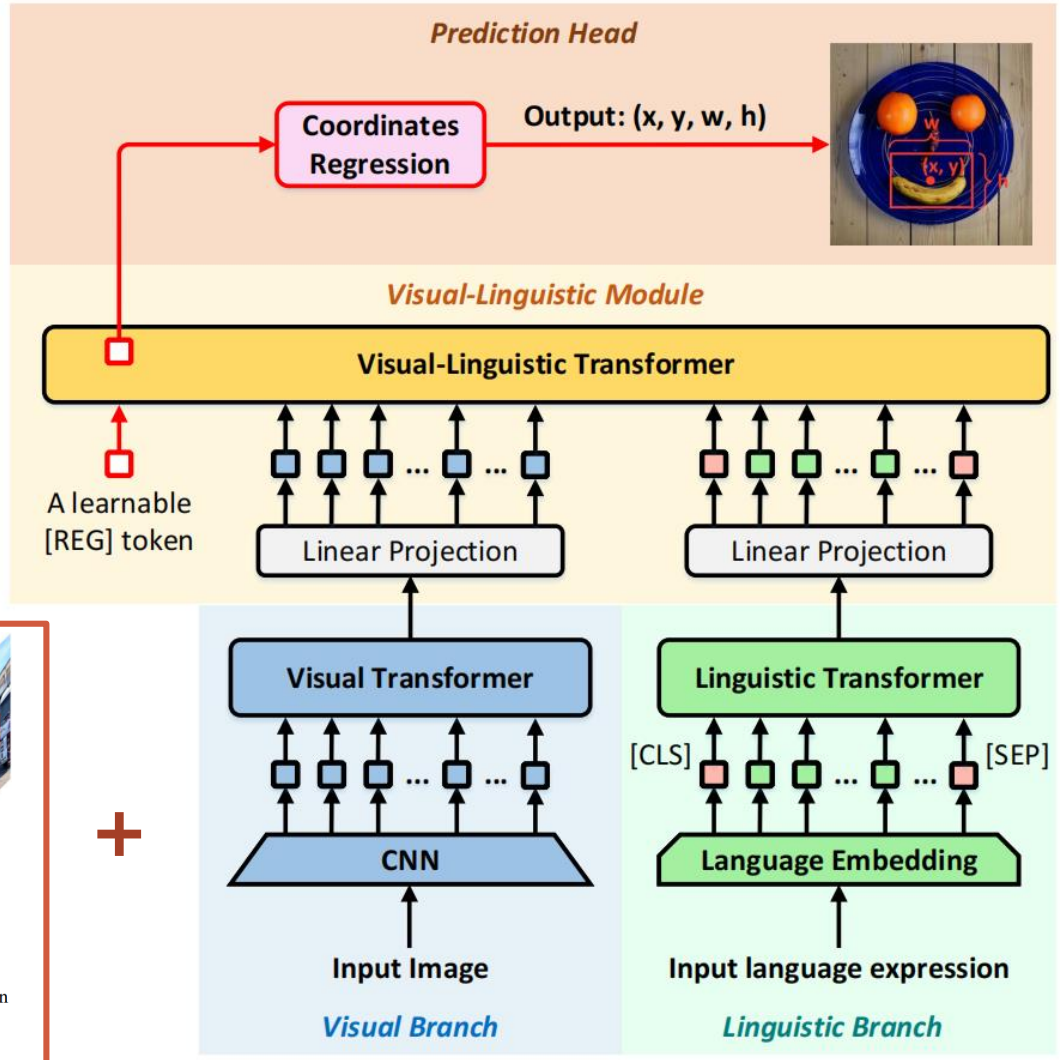
Referring Image Grounding

TransVG: End-to-End Visual Grounding with Transformers

Jiajun Deng[†], Zhengyuan Yang[‡], Tianlang Chen[‡], Wengang Zhou[†,§], Houqiang Li[†,§]
[†] CAS Key Laboratory of GIPAS, University of Science and Technology of China, Hefei, China
[‡] University of Rochester
[§] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
dengjj@mail.ustc.edu.cn



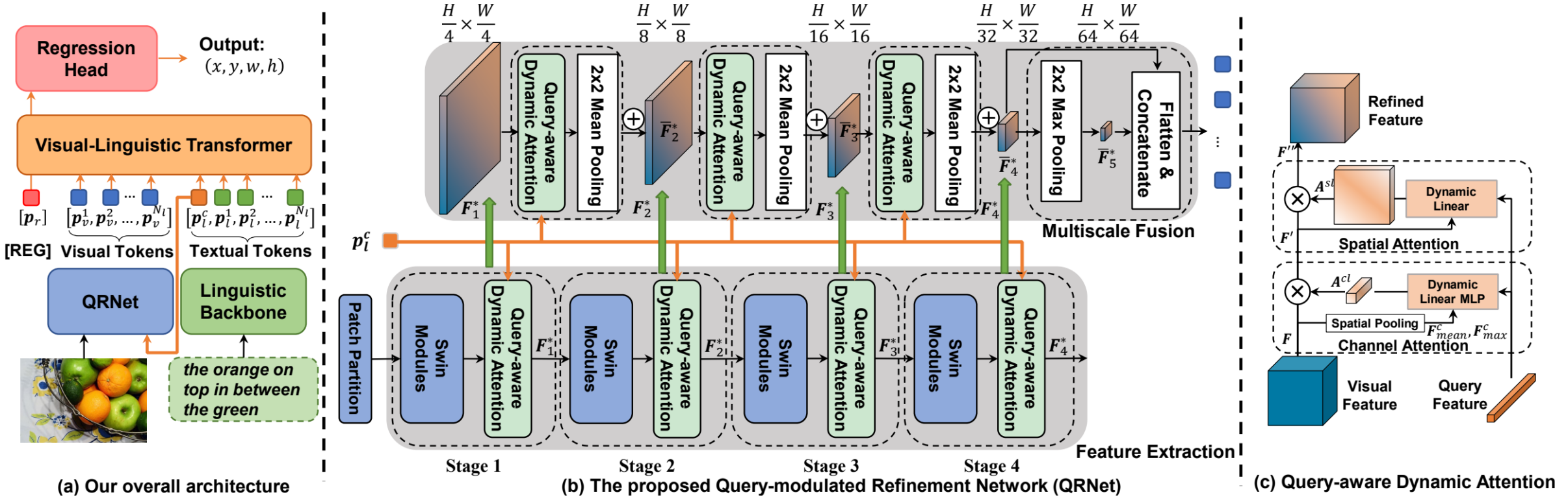
+



Referring Image Grounding

Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding

Jiabo Ye[1] Junfeng Tian[2] Ming Yan[2] Xiaoshan Yang[3]
Xuwu Wang[4] Ji Zhang[2] Liang He[1] Xin Lin[1]
[1]East China Normal University, Shanghai, China [2]Alibaba Group, Hangzhou, China
[3]NLPR, CASIA, Beijing, China [4]Fudan University, Shanghai, China



3/ Other Referring-related Tasks

- **Unsupervised VG**
  - (CVPR22) Generating Pseudo Language Queries for Visual Grounding
- **Referring Image Matting**
  - (NIPS22) Referring Image Matting

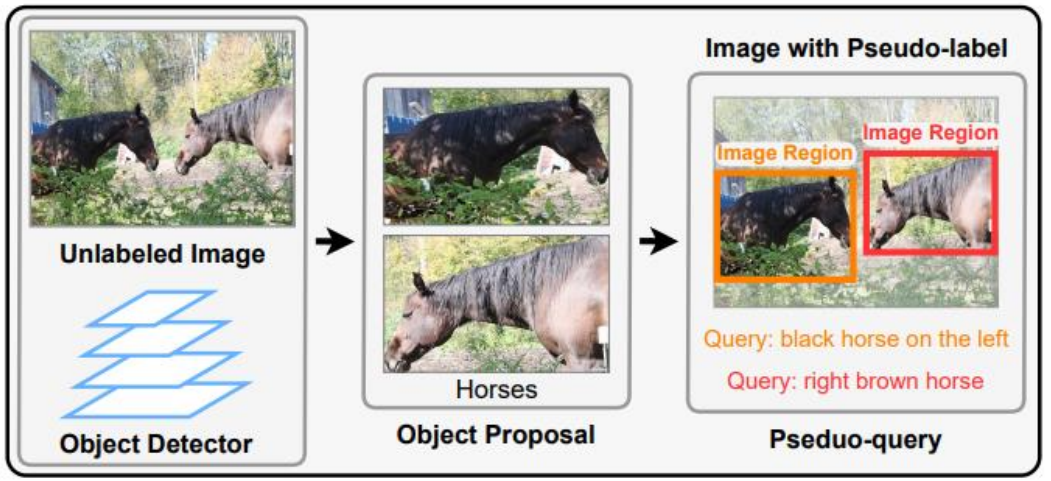
Unsupervised Referring Grounding

Pseudo-Q: Generating Pseudo Language Queries for Visual Grounding

Haojun Jiang[1]* Yuanze Lin[3]*† Dongchen Han[1] Shiji Song[1] Gao Huang[1,2]‡
[1]Tsinghua University, BNRist [2]BAAI [3]University of Washington
{jhj20, hdc19}@mails.tsinghua.edu.cn, yuanze@uw.edu,
{shijis, gaohuang}@tsinghua.edu.cn

(a) Fully-supervised VG

(b) Weakly-supervised VG



(c) Unsupervised VG

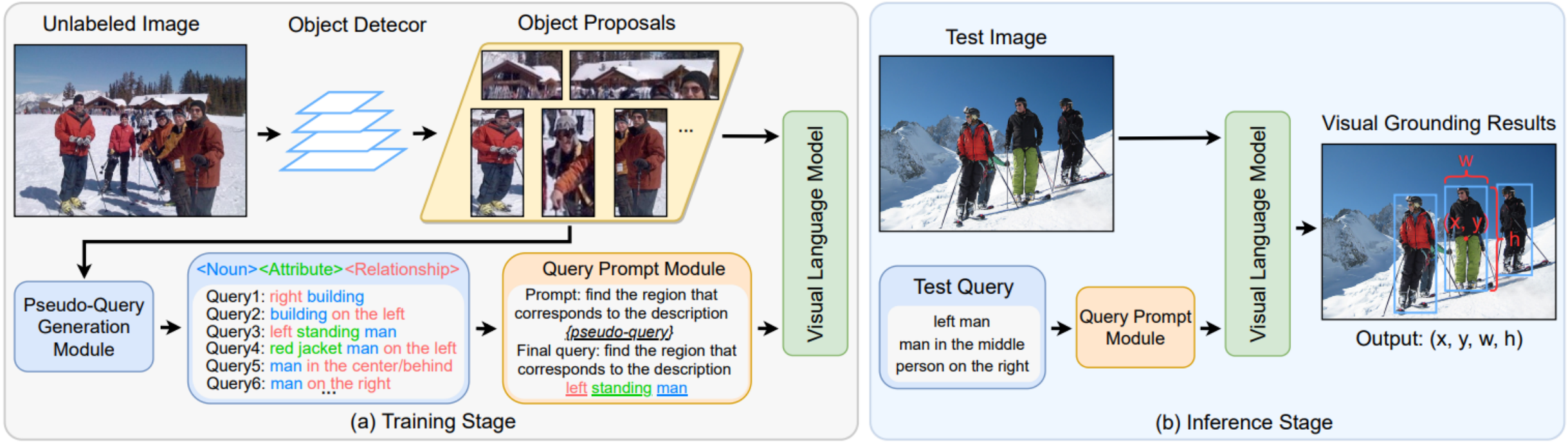


Figure 2. **Overview of our Pseudo-Q method. Better view in color and zoom in.** The proposed approach consists of a pseudo-query generation module, a query prompt module, and a visual-language model. (a) During the training stage, pseudo image region-query pairs are generated to train visual language model. (b) During the inference stage, the test query is filled into the prompt template, and the target object is located by the trained model.

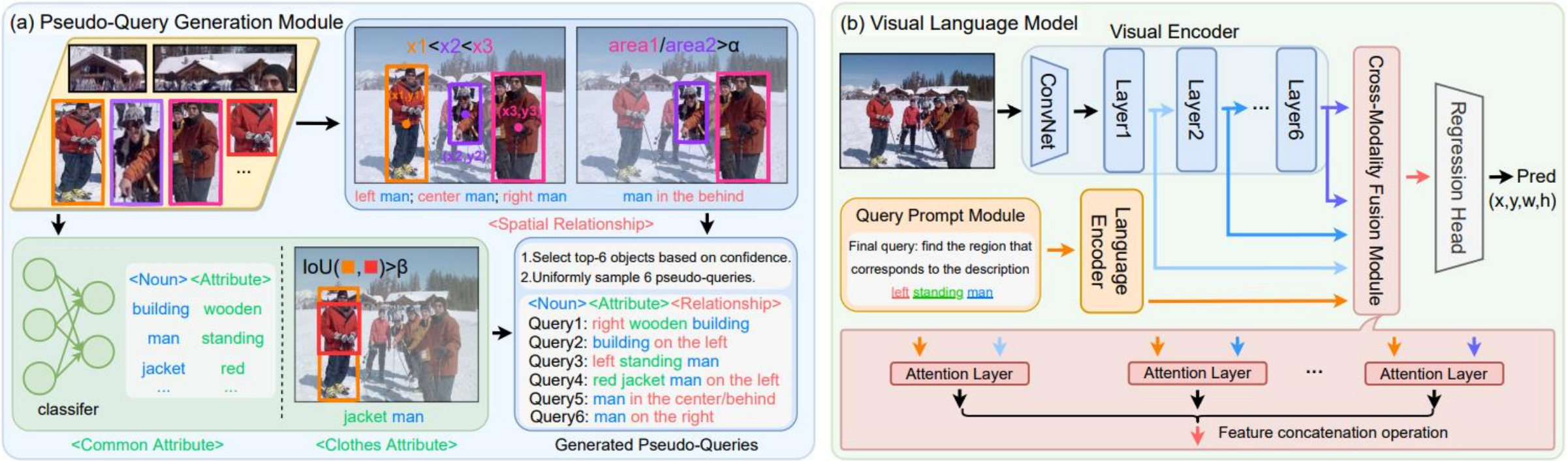


Figure 3. (a) The **pseudo-query generation module** produces spatial relationships and attributes for corresponding objects. (b) The **visual-language model** consists of a visual encoder, a language encoder, and a cross-modality fusion module.

- **Nouns:** off-the-shelf object detector
- **Attribute Classifier**
  - *Bottom-up and top-down attention for image captioning and visual question answering.*
- **Relationships**
  - 1. Horizontal(left, middle and right) 2. Vertical(top and bottom) 3. depth(front and behind)

Table 4. Pseudo-query templates. *Attr* and *Rela* represents attribute and relationship, respectively.

Pseudo Query Template	Example
{*Noun*}	"man", "building" etc.
{*Noun*} {*Attr*} {*Attr*} {*Noun*}	"man standing" etc. "talk man", "wooden building" etc.
{*Noun*} {*Rela*} {*Rela*} {*Noun*}	"man on the right" etc. "center man", "left building" etc.
{*Noun*} {*Attr*} {*Rela*} {*Noun*} {*Rela*} {*Attr*} {*Attr*} {*Noun*} {*Rela*} {*Attr*} {*Rela*} {*Noun*} {*Rela*} {*Noun*} {*Attr*} {*Rela*} {*Attr*} {*Noun*}	"man standing on the right" etc. "man right standing" etc. "standing man on the right" etc. "standing right man" etc. "right man standing" etc. "right standing man" etc.

Table 1. Comparison with state-of-the-art methods on RefCOCO [65], RefCOCO+ [65] and RefCOCOg [40] datasets in terms of top-1 accuracy (%). *"Sup."* refers to supervision level: No (without annotation), Weak (only annotated queries), Full (annotated bbox-query pairs). The best two results with supervision level of No and Weak are **bold-faced** and underlined, respectively.

Method	Published on	Sup.	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	testB	val	testA	testB	val-g	val-u	test-u
CPT [62]	*arXiv'21*	No	32.20	36.10	30.30	31.90	35.20	28.80	-	36.70	36.50
Ours	*CVPR'22*	No	**56.02**	**58.25**	**54.13**	38.88	**45.06**	32.13	**49.82**	**46.25**	**47.44**
VC [68]	*CVPR'18*	Weak	-	33.29	30.13	-	34.60	31.58	33.79	-	-
ARN [36]	*ICCV'19*	Weak	34.26	36.43	33.07	34.53	36.01	33.75	33.75	-	-
KPRN [37]	*ACMMM'19*	Weak	35.04	34.74	36.98	35.96	35.24	36.96	33.56	-	-
DTWREG [49]	*TPAMI'21*	Weak	39.21	41.14	37.72	**39.18**	40.10	**38.08**	43.24	-	-
MAttNet [64]	*CVPR'18*	Full	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
NMTree [35]	*ICCV'19*	Full	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
FAOA [61]	*ICCV'19*	Full	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
ReSC [60]	*ECCV'20*	Full	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
TransVG [13]	*ICCV'21*	Full	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44

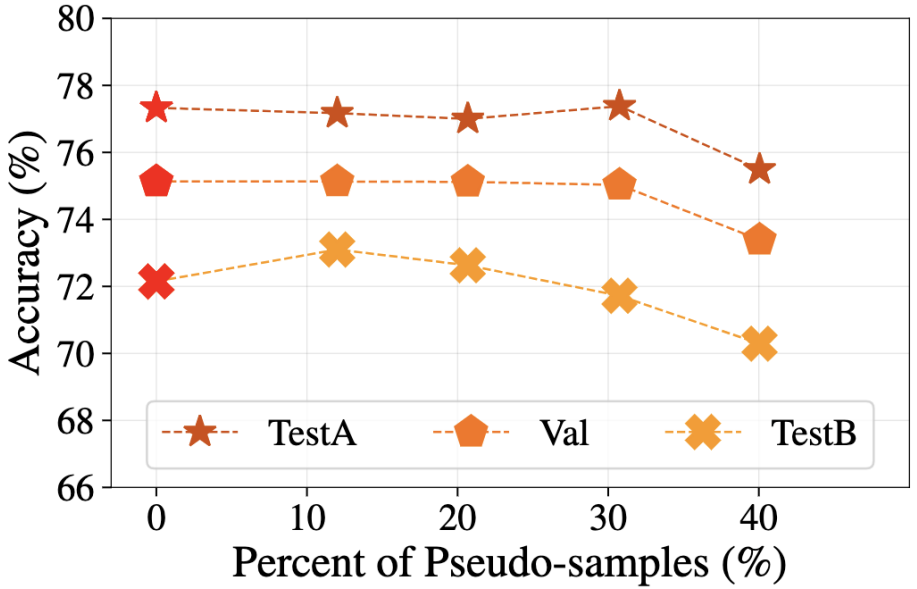


Figure 4. Experiments of reducing the manual labeling cost on RefCOCO [65]. We replace the manual labels whose queries contain spatial relationships with our pseudo-samples.

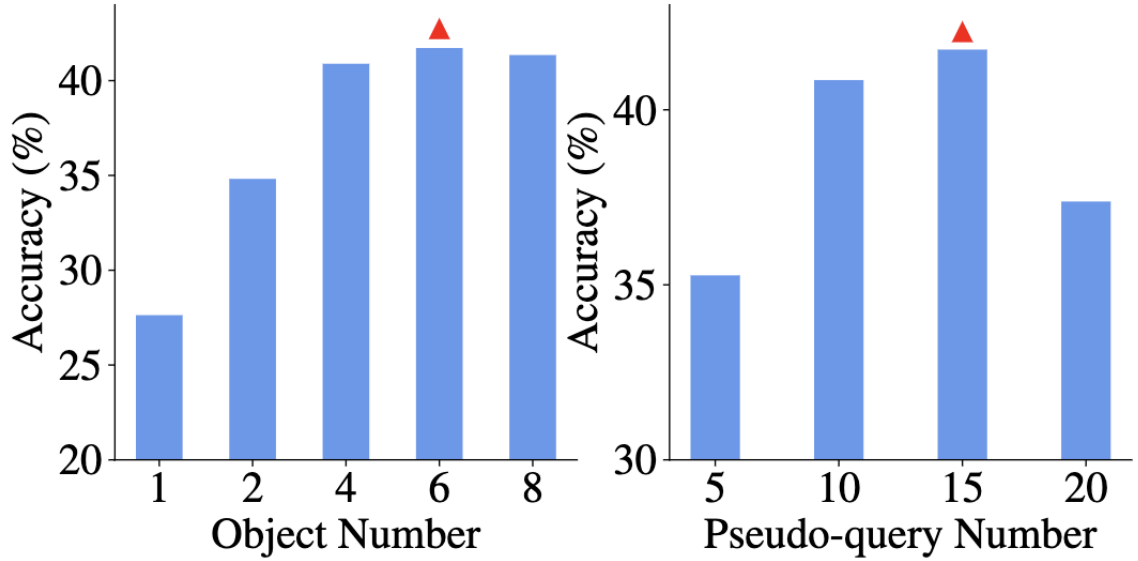


Figure 5. Left: Ablation of object number. Right: Ablation of pseudo-query number. Both are conducted on ReferItGame [40].

Noun	Attr	Rela	ML-CMA	Prompt	RefCOCO	ReferIt
✓					22.04	27.91
✓	✓				31.30 (↑9.26)	31.33 (↑3.42)
✓		✓			48.71 (↑**26.67**)	39.26 (↑**11.35**)
✓	✓	✓			53.39 (↑31.35)	40.37 (↑12.46)
✓	✓	✓	✓		55.16 (↑1.77)	41.72 (↑1.35)
✓	✓	✓	✓	✓	56.02 (↑2.63)	43.32 (↑2.95)

Referring Image Matting

Referring Image Matting

Jizhizi Li[1] **Jing Zhang**[1] **Dacheng Tao**[2,1]

[1]The University of Sydney, Australia,
[2]JD Explore Academy, China

Composition relation: left/right
Prompt: flower
Basic expression:
the lightpink and salient flower
Absolute position expression:
the plant which is lightpink and salient at the rightmost edge of the picture
Relative position expression:
the flower which is lightpink at the right side of the cat which is dimgray and non-transparent

(a)

Composition relation: top/bottom
Prompt: human
Basic expression:
the person in the linen lace
Absolute position expression:
the female human with the lightgray lace on top of the image
Relative position expression:
the female individual with the lightgray print over the male mortal who is dressed in white sleeve

(b)

Composition relation: in front of/behind
Prompt: alpaca
Basic expression:
the beast which is sienna and salient
Absolute position expression:
the alpaca which is sienna and non-transparent in front of the photo
Relative position expression:
the darksalmon and salient brute in front of the salient male mankind with the white t-shirt

(c)

Referring Image Matting

- **Pre-processing and filtering**
  - AM-2k, P3M-10k, and AIM-500
- **Annotate the category names of entities**
  - Mask RCNN to detect and manually check and correct
  - Adopt WordNet to generate synonyms and manually check and replace with more reasonable ones
- **Annotate the attributes of entities**
  - **Color:** cluster all the pixel values of the foreground image, find out the most frequent value, and match it with the specific color in webcolors
  - **Gender, age:** *Age and gender classification using convolutional neural networks. 2015 (CVPRW)*
  - **Clothes type:** *Mmfashion: An open-source toolbox for visual fashion analysis. In ACM Multimedia 2021, Open Source Software Competition, 2021*
  - **Salient/Transparent or not:** *Deep automatic natural image matting. IJCAI-21*

Referring Image Matting

Expression generation engine To provide abundant expressions for the entities in the composite images, we define three types of expressions for each entity from the aspect of different logic forms defined as follows, where <att_i> stands for the attribute, <obj_i> stands for the category name, and <rel_i> stands for the relationship between the reference entity and the related one:

1. ***Basic expression*** This is the expression that describes the target entity with as many attributes as one can, e.g, `the/a` <att_0> <att_1>...<obj_0> or `the/a` <obj_0> `which/that is` <att_0> <att_1>, `and` <att_2>. For example, as shown in Figure 3(a), the basic expression for the flower entity is '`the lightpink and salient flower`';
2. ***Absolute position expression*** This is the expression that describes the target entity with many attributes and its absolute position in the image, e.g., `the/a` <att_0> <att_1>...<obj_0> <rel_0> `the photo/image/picture` or `the/a` <obj_0> `which/that is` <att_0> <att_1> <rel_0> `the photo/image/picture`. For example, as shown in Figure 3(a), the absolute position expression for the flower is '`the plant which is lightpink and salient at the rightmost edge of the picture`';
3. ***Relative position expression*** This is the expression that describes the target entity with many attributes and its relative position with another entity, e.g., `the/a` <att_0> <att_1>...<obj_0> <rel_0> `the/a` <att_2> <att_3>...<obj_1> or `the/a` <obj_0> `which/that is` <att_0> <att_1> <rel_0> `the/a` <obj_1> `which/that is` <att_2> <att_3>. For example, as shown in Figure 3(a), the relative position expression for the flower is '`the flower which is lightpink at the right side of the cat which is dimgray and non-transparent`'.

Referring Image Matting

Table 1: Statistics of RefMatte and RefMatte-RW100 in regarding to the number of images, alpha mattes, text descriptions, categories, attributes, relationship words, and average length of texts.

Dataset	Setting	Split	Images Num.	Mattes Num.	Text Num.	Categories Num.	Attrs. Num.	Rels. Num.	Texts Length.
RefMatte	Prompt	train	30,391	77,849	77,849	230	-	-	1.06
		test	1,602	4,085	4,085	66	-	-	1.04
	Expression	train	45,000	112,506	449,624	230	132	31	16.86
		test	2,500	6,243	24,972	66	102	31	16.80
RefMatte-RW100	Expression	test	100	221	884	29	135	34	12.01

(a) (b) (c)

Figure 4: The wordcloud of the prompts (a), attributes (b), and relationships (c) in RefMatte.

Summary

- **There is already good performance on classical tasks:**
  - Referring Image Grounding
  - Referring Image Segmentation
  - Referring Video Segmentation
- **Model frameworks tend to be similar**
- **Explore other settings like unsupervised**
- **Explore language-driven other vision tasks like matting/editing/generation**

Examples from “Blended Diffusion for Text-Driven Editing of Natural Images”

Thanks