ZHEJIANG UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

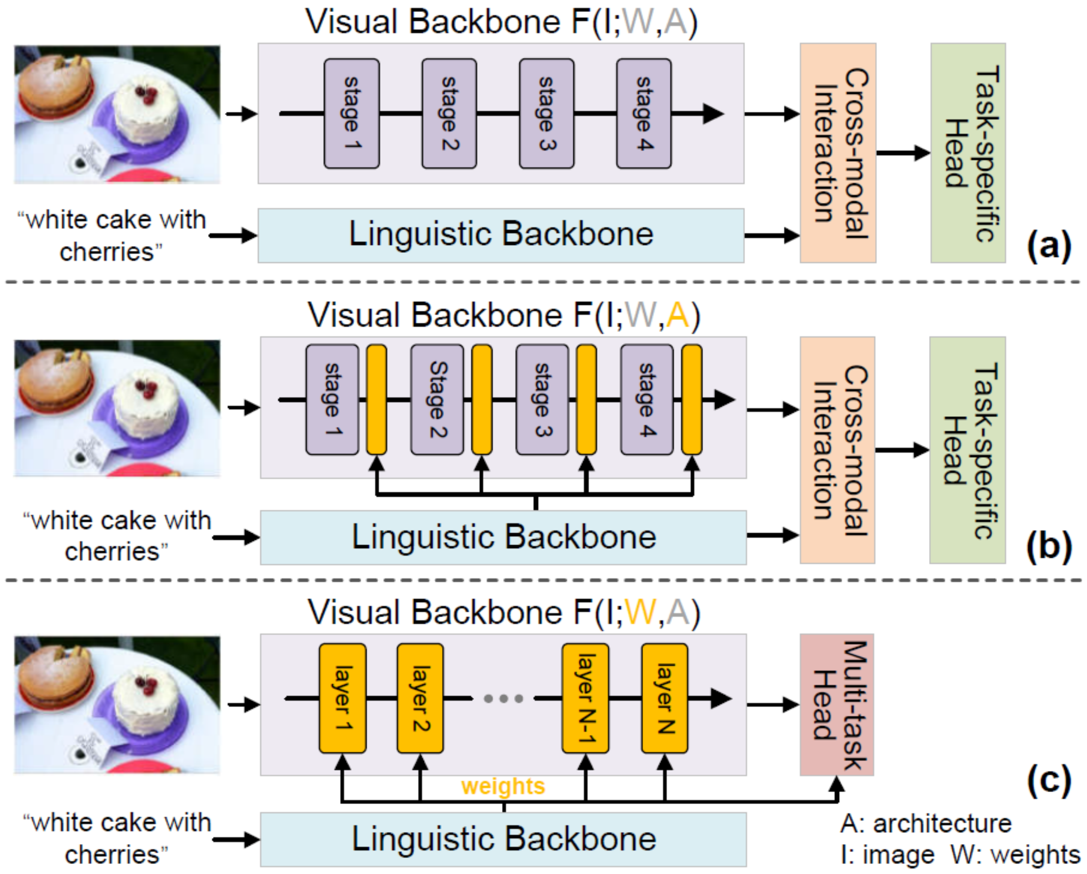
Language Adaptive Weight Generation for Multi-task Visual Grounding

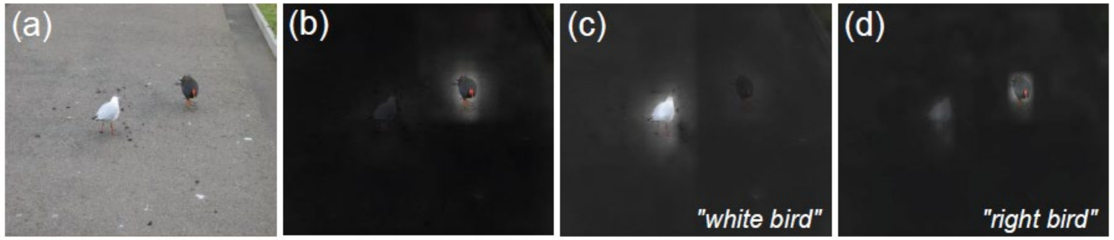
Wei Su[1], Peihan Miao[1], Huanzhang Dou[1], Gaoang Wang[4]
Liang Qiao[1,3], Zheyang Li[1,3], Xi Li[1,2,5]

[1]Zhejiang University [2]Shanghai AI Laboratory [3]Hikvision Research Institute
[4]Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University
[5]Shanghai Institute for Advanced Study of Zhejiang University

Tag: WED-AM-250

Motivation





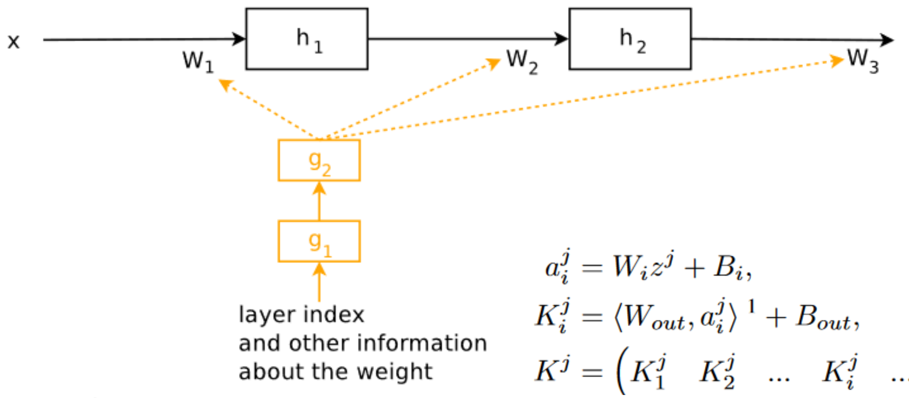
- ➢ The visual backbone passively extracts features
- ➢ Fixed architectures and weights
- ➢ Regardless of referring expressions
- ➢ Mismatches between the extracted visual features and those required for various referring expressions
- ➢ Missing or redundant visual features

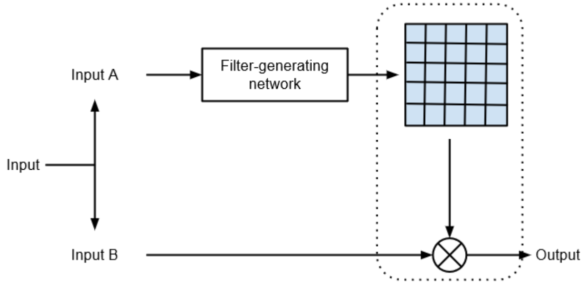
✓ An active perception visual grounding framework, which can actively extract expression-relevant visual features.

Related Work: Dynamic Weight Networks

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- **Direct Generation (FCs):**
  Hypernetworks
  Dynamic Filter Networks
- **Weighted Summation:**
  Dynamic Convolution
  Conditional Convolution
  Omni-dimensional Convolution
- **Matrix Decomposition:**
  Dynamic Channel Fusion

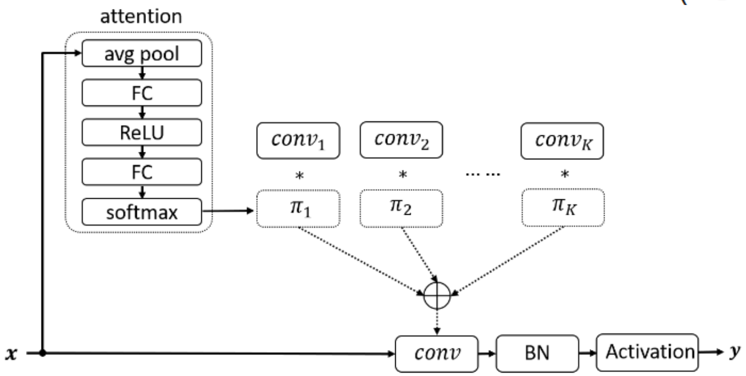




$$a_i^j = W_i z^j + B_i, \quad \forall i = 1, .., N_{in}, \forall j = 1, ..., D$$

$$K_i^j = \langle W_{out}, a_i^j \rangle^{1} + B_{out}, \quad \forall i = 1, .., N_{in}, \forall j = 1, ..., D$$

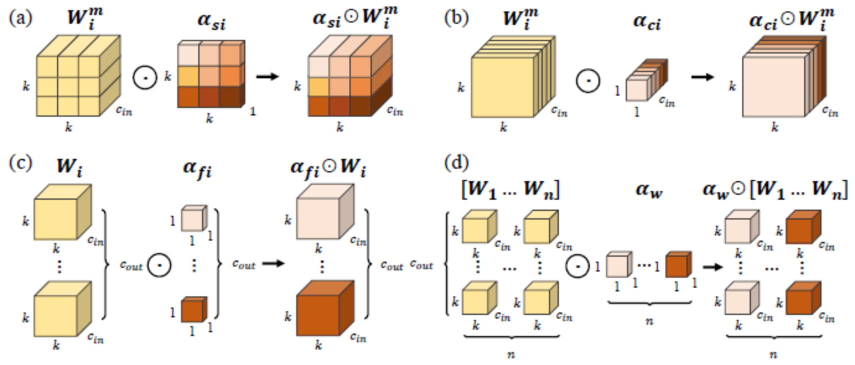
$$K^j = \left(K_1^j \quad K_2^j \quad \dots \quad K_i^j \quad \dots \quad K_{N_{in}}^j \right), \quad \forall j = 1, ..., D$$



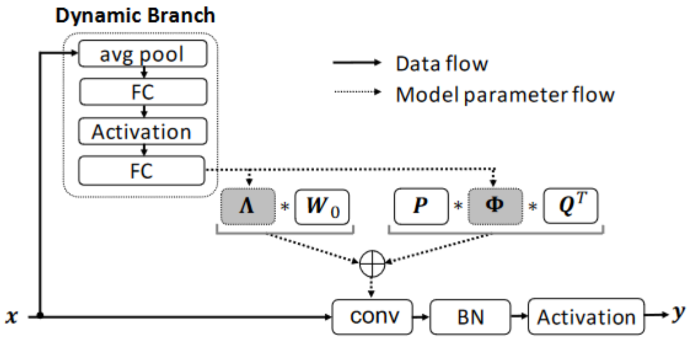
$$y = g(\tilde{W}^T(x)x + \tilde{b}(x))$$

$$\tilde{W}(x) = \sum_{k=1}^{K} \pi_k(x)\tilde{W}_k, \ \tilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\tilde{b}_k$$

$$\text{s.t.} \quad 0 \le \pi_k(x) \le 1, \sum_{k=1}^{K} \pi_k(x) = 1,$$



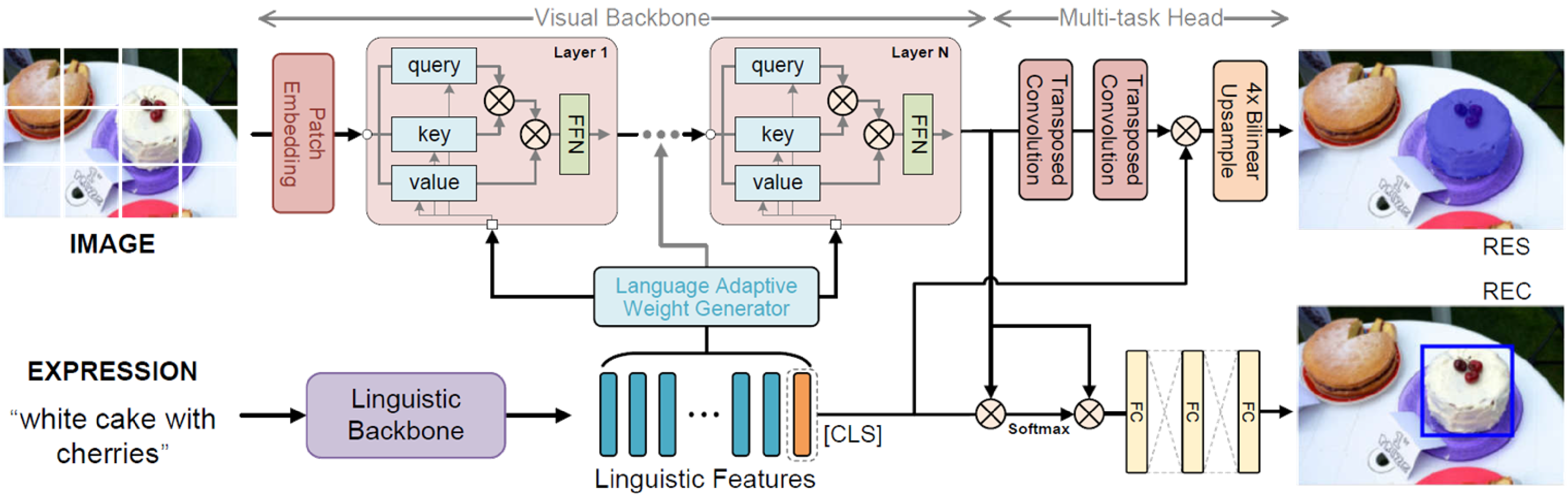
$$y = (\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \dots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n) * x$$



$$W(x) = W_0 + P\Phi(x)Q^T = W_0 + \sum_{i=1}^{L}\sum_{j=1}^{L} p_i \phi_{i,j}(x) q_j^T$$

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

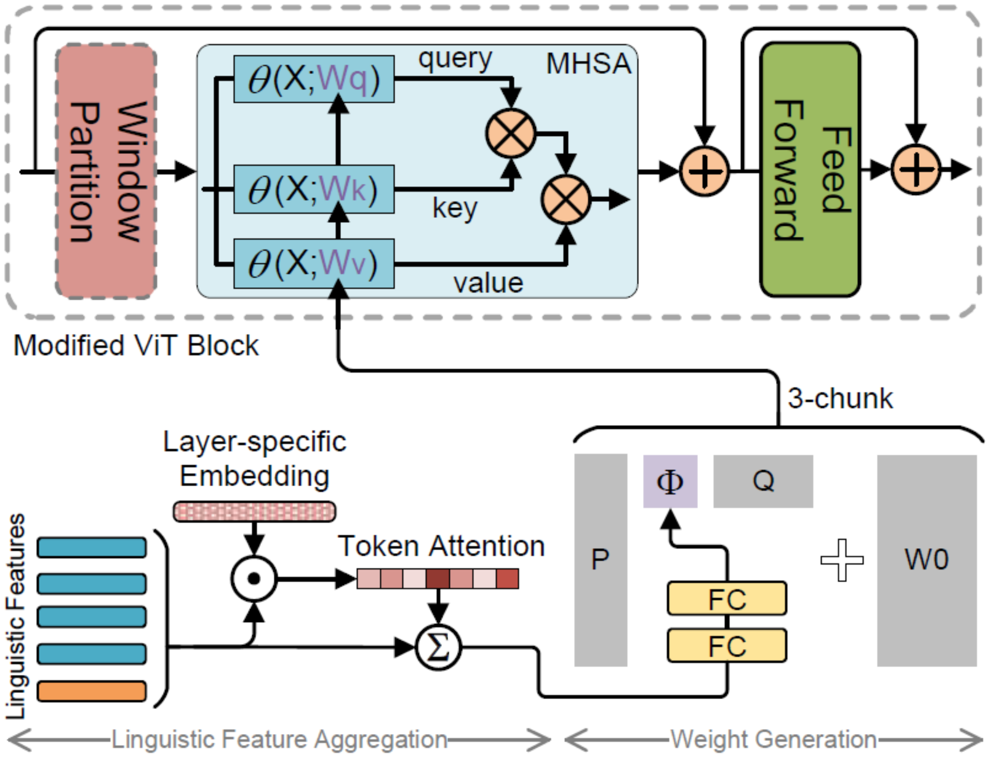
Method: VG-LAW



An active perception framework for multi-task visual grounding based on the language adaptive weights

- Components: visual backbone, linguistic backbone, multi-task head, language adaptive weight generator
- Language adaptive weights: mapping linguistic features to weights of the visual backbone
- Visual backbone can actively extract expression-relevant visual features using language-adaptive weights
- Don't need to modify the visual backbone architecture or elaborately design cross-modal interaction modules

Method: Language Adaptive Weight Generation



- **Linguistic Feature Aggregation**

  Aggregate linguistic features with fixed sizes for each layer independently.

$$\alpha_i^g = \mathrm{Softmax}([e_i^g \cdot F_l^{g,1}, e_i^g \cdot F_l^{g,2}, \cdots, e_i^g \cdot F_l^{g,L}])$$

$$h_0^{i,g} = \sum_{j=1}^{L} \alpha_i^{g,j} F_l^{g,j}, \quad h_1^i = \delta(W_1^i h_0^i)$$

- **Weight Generation**

  Map linguistic features to language-adaptive weights using the multi-head attention mechanism

$$[W_q^i, W_k^i, W_v^i] = W_0^i + P\Phi(h_1^i)Q^T$$

$$X_q = \theta(X; W_q), \; X_k = \theta(X; W_k), \; X_v = \theta(X; W_v)$$

Experiments

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Training Objectives

- REC: L1 loss and Generalized IoU loss
- RES: Focal loss and DICE/F-1 loss

$$\mathcal{L}_{\text{det}} = \lambda_{L1}\mathcal{L}_{L1}(b,\hat{b}) + \lambda_{\text{giou}}\,\mathcal{L}_{\text{giou}}\,(b,\hat{b})$$

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{focal}}\,\mathcal{L}_{\text{focal}}\,(s,\hat{s}) + \lambda_{\text{dice}}\,\mathcal{L}_{\text{dice}}\,(s,\hat{s})$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{seg}}$$

Configs

- Datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame
- Evaluation Protocol: Prec@0.5 for REC and mIoU for RES

Experimental Results

➢ Comparison with State-of-the-art REC Methods

Methods	Venue	Visual Backbone	Multi-task	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test	ReferItGame test
Two-stage:												
MAttNet [46]	CVPR18	RN101	✗	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27	29.04
RvG-Tree [13]	TPAMI19	RN101	✗	75.06	78.61	69.85	63.51	67.45	56.66	66.95	66.51	-
CM-A-E [30]	CVPR19	RN101	✗	78.35	83.14	71.32	68.09	73.65	58.03	67.99	68.67	-
Ref-NMS [2]	AAAI21	RN101	✗	80.70	84.00	76.04	68.25	73.68	59.42	70.55	70.62	-
One-stage:												
FAOA [43]	ICCV19	DN53	✗	72.54	74.35	68.50	56.81	60.23	49.60	61.33	60.36	60.67
ReSC-Large [42]	ECCV20	DN53	✗	77.63	80.45	72.30	63.59	68.36	56.81	67.30	67.20	64.60
MCN [33]	CVPR20	DN53	✓	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01	-
RealGIN [49]	TNNLS21	DN53	✗	77.25	78.70	72.10	62.78	67.17	54.21	62.75	62.33	-
PLV-FPN* [26]	TIP22	RN101	✗	81.93	84.99	76.25	71.20	77.40	61.08	70.45	71.08	71.77
Transformer-based:												
TransVG [4]	ICCV21	RN101	✗	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	70.73
RefTR* [23]	NeurIPS21	RN101	✓	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40	71.42
SeqTR [50]	ECCV22	DN53	✗	81.23	85.00	76.08	68.82	75.37	58.78	71.35	71.58	69.66
Word2Pix [48]	TNNLS22	RN101	✗	81.20	84.39	78.12	69.74	76.11	61.24	70.81	71.34	-
YORO [12]	ECCVW22	-	✗	82.90	85.60	77.40	73.50	78.60	64.90	73.40	74.30	71.90
QRNet [45]	CVPR22	Swin-S	✗	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03	74.61
Ours:												
VG-LAW	-	ViT-B	✗	86.06	88.56	82.87	75.74	80.32	66.69	75.31	75.95	76.60
VG-LAW	-	ViT-B	✓	86.62	89.32	83.16	76.37	81.04	67.50	76.90	76.96	77.22

Comparisons on the RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame datasets. RN101, DN53, Swin-S, and ViT-B are shorthand for the ResNet101, DarkNet53, Swin-Transformer Small, and ViT Base, respectively. We highlight the best and second-best performance in the red and blue colors.

Experimental Results

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

➢ **Comparison with State-of-the-art RES Methods**

Methods	Venue	Visual Backbone	Multi-task	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
CGAN [32]	MM20	DN53	✗	64.86	68.04	62.07	51.03	55.51	44.06	54.40	54.25
MCN [33]	CVPR20	DN53	✓	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
LTS [17]	CVPR21	DN53	✗	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [50]	ICCV21	DN53	✗	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
RefTR* [23]	NeurIPS21	RN101	✓	70.56	73.49	66.57	61.08	64.69	52.73	58.73	58.51
SeqTR [50]	ECCV22	DN53	✗	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
LAVT* [44]	CVPR22	Swin-B	✗	74.46	76.89	70.94	65.81	70.97	59.23	63.62	63.66
Ours:											
VG-LAW	-	ViT-B	✗	75.05	77.36	71.69	66.61	70.30	58.14	65.36	65.13
VG-LAW	-	ViT-B	✓	75.62	77.51	72.89	66.63	70.38	58.89	65.63	66.08

Comparisons on the RefCOCO, RefCOCO+, and RefCOCOg for RES task. * represents ImageNet pre-training. RN101, DN53, Swin-B, and ViT-B are shorthand for the ResNet101, DarkNet53, Swin-Transformer Base, and ViT Base, respectively. We highlight the best and second-best performance in the red and blue colors.

Experimental Results

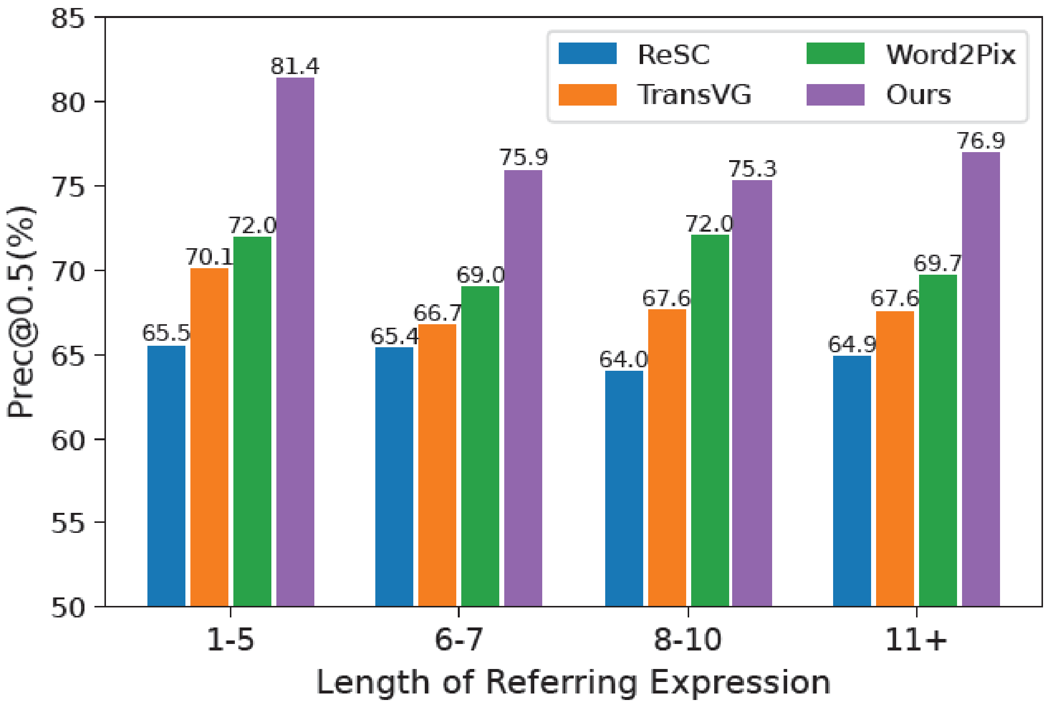
➢ **Ablation Analysis**

LAWG	LAP	MTH	*Prec@0.5(%)*
✓			74.89
	✓		74.37
✓	✓		76.60
✓	✓	✓	77.22

Ablation experiments on ReferItGame to evaluate the proposed language adaptive weight generation (**LAWG**), language adaptive pooling (**LAP**), and multi-task head (**MTH**).

Experimental Results

➢ **Analysis of Referring Expression Length**



No significant performance degradation when the length of referring expressions varies from 6-7 to 11+.

Experimental Results

➢ **Qualitative results**

From left to right: the input image, the ground truth of REC and RES, the prediction of VG-LAW, and the attention of the visual backbone with language-adaptive weights.

Conclusion

Contributions

- ✓ An active perception visual grounding framework based on the language adaptive weights (VG-LAW), actively extracts the expression-relevant visual features.
- ✓ A neat yet efficient multi-task head for REC and RES tasks jointly without carefully designed cross-modal interaction modules.
- ✓ State-of-the-art performance on RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame.

Future Works

- ☐ Search for the suitable positions of modules to inject the linguistic information/priors, instead of modifying the weights of all modules in the visual backbone.
- ☐ Enhance the multi-task head to enable multi-instance referring detection and segmentation.
- ☐ Expand VG-LAW to more types of visual backbones, e.g. Resnet series.