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
{weisuzju, peihan.miao, hzdou, qiaoliang, xilizju}@zju.edu.cn
gaoangwang@intl.zju.edu.cn, lizheyang@hikvision.com

Abstract

Although the impressive performance in visual grounding, the prevailing approaches usually exploit the visual backbone in a passive way, i.e., the visual backbone extracts features with fixed weights without expression-related hints. The passive perception may lead to mismatches (e.g., redundant and missing), limiting further performance improvement. Ideally, the visual backbone should actively extract visual features since the expressions already provide the blueprint of desired visual features. The active perception can take expressions as priors to extract relevant visual features, which can effectively alleviate the mismatches. Inspired by this, we propose an active perception ***V****isual* ***G****rounding framework based on* ***L****anguage* ***A****daptive* ***W****eights, called VG-LAW. The visual backbone serves as an expression-specific feature extractor through dynamic weights generated for various expressions. Benefiting from the specific and relevant visual features extracted from the language-aware visual backbone, VG-LAW does not require additional modules for cross-modal interaction. Along with a neat multi-task head, VG-LAW can be competent in referring expression comprehension and segmentation jointly. Extensive experiments on four representative datasets, i.e., RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame, validate the effectiveness of the proposed framework and demonstrate state-of-the-art performance.*

1. Introduction

Visual grounding (such as referring expression comprehension [4, 23, 42, 45, 46, 48, 50], referring expression segmentation [6, 14, 17, 23, 32, 33, 44], and phrase grounding [4, 23, 50]) aims to detect or segment the specific object based on a given natural language description. Compared to general object detection [38] or instance segmentation [11], which can only locate objects within a predefined and fixed category set, visual grounding is more flexible and purposeful. Free-formed language descriptions can specify specific visual properties of the target object, such as categories, attributes, relationships with other objects, relative/absolute positions, and *etc*.

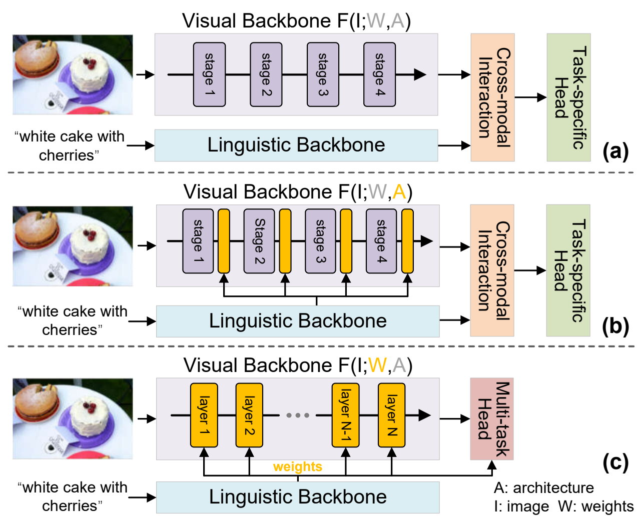

Figure 1. The comparison of visual grounding frameworks. (a) The visual and linguistic backbone independently extracts features, which are fused through cross-modal interaction. (b) Additional designed modules are inserted into the visual backbone to modulate visual features using linguistic features. (c) VG-LAW can generate language-adaptive weights for the visual backbone and directly output referred objects through our designed multi-task head without additional cross-modal interaction modules.

Due to the similarity with detection tasks, previous visual grounding approaches [23, 33, 46, 50] usually follow the general object detection frameworks [1, 11, 37], and pay

*corresponding author.

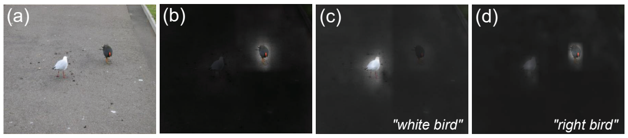

Figure 2. Attention visualization of the visual backbone with different weights. (a) input image, (b) visual backbone with fixed weights, (c) and (d) visual backbone with weights generated for "white bird" and "right bird", respectively.

attention to the design of cross-modal interaction modules. Despite achieving impressive performance, the visual backbone is not well explored. Concretely, the visual backbone passively extracts visual features with fixed architecture and weights, regardless of the referring expressions, as illustrated in Fig. 1 (a). Such passive feature extraction may lead to mismatches between the extracted visual features and those required for various referring expressions, such as missing or redundant features. Taking Fig. 2 as an example, the fixed visual backbone has an inherent preference for the image, as shown in Fig. 2 (b), which may be irrelevant to the referring expression "white bird". Ideally, the visual backbone should take full advantage of expressions, as the expressions can provide information and tendencies about the desired visual features.

Several methods have noticed this phenomenon and proposed corresponding solutions, such as QRNet [45], and LAVT [44]. Both methods achieve the expression-aware visual feature extraction by inserting carefully designed interaction modules (such as QD-ATT [45], and PWAN [44]) into the visual backbone, as illustrated in Fig. 1 (b). Concretely, visual features are first extracted and then adjusted using QD-ATT (channel and spatial attention) or PWAM (transformer-based pixel-word attention) in QRNet and LAVT at the end of each stage, respectively. Although performance improvement with adjusted visual features, the extract-then-adjust paradigm inevitably contains a large number of feature-extraction components with fixed weights, *e.g.,* the components belonging to the original visual backbone in QRNet and LAVT. Considering that the architecture and weights jointly determine the function of the visual backbone, this paper adopts a simpler and fine-grained scheme that modifies the function of the visual backbone with language-adaptive weights, as illustrated in Fig. 1 (c). Different from the extract-then-adjust paradigm used by QRNet and LAVT, the visual backbone equipped with language-adaptive weights can directly extract expression-relevant visual features without additional feature-adjustment modules.

In this paper, we propose an active perception **V**isual **G**rounding framework based on **L**anguage **A**daptive **W**eights, called VG-LAW. It can dynamically adjust the behavior of the visual backbone by injecting the information of referring expressions into the weights. Specifically, VG-LAW first obtains the specific language-adaptive weights for the visual backbone through two successive processes of linguistic feature aggregation and weight generation. Then, the language-aware visual backbone can extract expression-relevant visual features without manually modifying the visual backbone architecture. Since the extracted visual features are highly expression-relevant, cross-modal interaction modules are not required for further cross-modal fusion, and the entire network architecture is more streamlined. Furthermore, based on the expression-relevant features, we propose a lightweight but neat multi-task prediction head for jointly referring expression comprehension (REC) and referring expression segmentation (RES) tasks. Extensive experiments on RefCOCO [47], RefCOCO+ [47], RefCOCOg [36], and ReferItGame [19] datasets demonstrate the effectiveness of our method, which achieves state-of-the-art performance.

The main contributions can be summarized as follows:

- We propose an active perception visual grounding framework based on the language adaptive weights, called VG-LAW, which can actively extract expression-relevant visual features without manually modifying the visual backbone architecture.
- Benefiting from the active perception of visual feature extraction, we can directly utilize our proposed neat but efficient multi-task head for REC and RES tasks jointly without carefully designed cross-modal interaction modules.
- Extensive experiments demonstrate the effectiveness of our framework, which achieves state-of-the-art performance on four widely used datasets, *i.e.,* RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame.

2. Related Work

2.1. Referring Expression Comprehension

Referring expression comprehension (REC) [4,13,30,42,43,46,48–50] aims to generate a bounding box in an image specified by a given referring expression. Early researchers explore REC through a two-stage framework [13,29,30,46], where region proposals [38] are first extracted and then ranked according to their similarity scores with referring expressions. To alleviate the speed and accuracy issues of the region proposals in the two-stage framework, simpler and faster one-stage methods [42, 43, 49] based on dense anchors are proposed. Recently, transformer-based methods [4,12,18,48,50] can effectively capture intra- and inter-modality context and achieve better performance, benefiting from the self-attention mechanism [40].

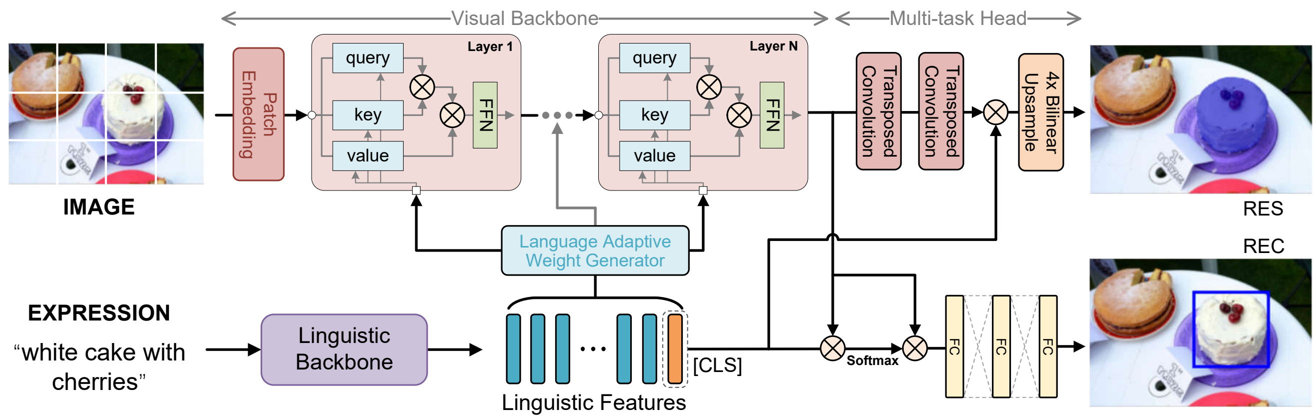

Figure 3. The overall architecture of our proposed VG-LAW framework. It consists of four components: (1) Linguistic Backbone, which extracts linguistic features from free-formed referring expressions, (2) Language Adaptive Weight Generator, which generates dynamic weights for the visual backbone conditioned on specific expressions, (3) Visual Backbone, which extracts visual features from the raw image and its behavior can be modified by language-adaptive weights, and (4) Multi-task Head, which predicts the bounding box and mask of referred object jointly. $\otimes$ represents the matrix multiplication.

2.2. Referring Expression Segmentation

Similar to REC, referring expression segmentation (RES) [6, 9, 14, 15, 17, 20, 23, 32, 44, 50] aims to predict a precise pixel-wise binary mask corresponding to the given referring expression. The pioneering work [14] proposes to generate segmentation masks for natural language expressions by concatenating the visual and linguistic features and mixing these two modal features with fully convolutional classifiers. Follow-up solutions [9, 15, 17, 32] propose various attention mechanisms to perform cross-modal interaction to generate a high-resolution segmentation map. Recent studies [6, 20, 23, 44, 50], like REC, leverage transformer [40] to realize cross-modal interaction and achieve excellent performance. All these methods achieve cross-modal interaction by either adjusting the inputs or modifying the architectures with fixed network weights.

2.3. Dynamic Weight Networks

Several works [3, 10, 16, 24, 41] have investigated dynamic weight networks, where given inputs adaptively generate the weights of the network. According to the way of dynamic weight generation, the current methods can be roughly divided into three categories. (1) Dynamic weights are directly generated using fully-connected layers with learnable embeddings [10] or intermediate features [16] as input. (2) Weights are computed as the weighted sum of a set of learnable weights [3, 22, 41], which can also be regarded as the mixture-of-experts and may suffer from challenging joint optimization. (3) The weights are analyzed from the perspective of matrix decomposition [24], and the final dynamic weights are generated by calculating the multiplication of several matrices.

3. Method

In this section, we will introduce the active perception framework for multi-task visual grounding, including the language-adaptive weight generation, multi-task prediction head, and training objectives.

3.1. Overview

The extraction of visual features by the visual backbone in the manner of passive perception may cause mismatch problems, which can lead to suboptimal performance despite subsequent carefully designed cross-modal interaction modules. Considering that expressions already provide a blueprint for the desired visual features, we propose an active perception visual grounding framework based on the language adaptive weights, called VG-LAW, as illustrated in Fig. 3. In this framework, the visual backbone can actively extract expression-relevant visual features using language-adaptive weights, without needing to manually modify the visual backbone architecture or elaborately design additional cross-modal interaction modules.

Specifically, the VG-LAW framework consists of four components, *i.e.,* linguistic backbone, language adaptive weight generator, visual backbone, and multi-task head. Given a referring expression, the N-layer BERT-based [5] linguistic backbone tokenizes the expressions, prepends a [CLS] token, and extracts linguistic features $F_l \in \mathbb{R}^{L\times d_l}$, where L and d_l represent the token numbers and dimension of linguistic features, respectively. The linguistic features F_l are then fed to the language adaptive weight generator to generate weights for the transformer-based visual backbone. Next, given an image $I \in \mathbb{R}^{3\times H\times W}$, the expression-aware visual features $F_v \in \mathbb{R}^{C\times \frac{H}{s}\times \frac{W}{s}}$ can be extracted by

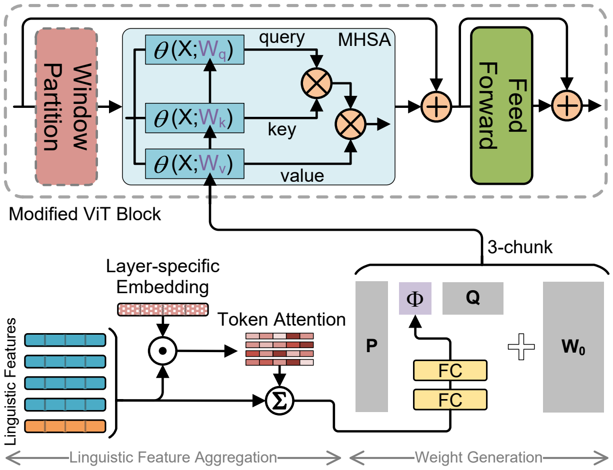

Figure 4. The detailed architecture for language adaptive weight generation. The upper part shows the architecture of the adapted ViT block in the visual backbone, and the lower part shows the linguistic feature aggregation and weight generation.

the visual backbone, where C and s represent the channel number and stride of the visual features, respectively. Finally, we pass the linguistic features $F_l^1 \in \mathbb{R}^{d_l}$ represented by the [CLS] token and the visual features to the multi-task head, which predicts the bounding box and mask of the referred object for REC and RES, respectively.

3.2. Language Adaptive Weight Generation

After extracting linguistic features, language-adaptive weights are generated to guide the active perception of the visual backbone. The process of language adaptive weight generation has two stages, *i.e.,* the layer-wise linguistic feature aggregation and the weight generation.

Linguistic Feature Aggregation. Considering the referring expressions correspond to a different number of linguistic tokens and each layer of the visual backbone may prefer different linguistic tokens, we try to aggregate linguistic features with fixed sizes for each layer independently. Inspired by the multi-head attention mechanism [40], we introduce a learnable layer-specific embedding $e_i \in \mathbb{R}^{d_l}$ for each layer i of the visual backbone to extract layer-specific linguistic features dynamically, which can improve the model flexibility at negligible cost. The calculation is performed on G groups. For each group g, the token-wise attention $\alpha_i^g \in [0,1]^L$ is assigned to the normalized dot product of e_i^g and F_l^g, which is denoted as:

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

Then, the aggregated linguistic feature $h_0^i \in \mathbb{R}^{d_l}$ can be derived by concatenating $h_0^{i,g} = \sum_{j=1}^{L} \alpha_i^{g,j} F_l^{g,j}$.

Finally, we use a fully-connected layer (FC) to reduce the dimension of the aggregated linguistic features for the i-th layer of the visual backbone, which is indicated as:

$$h_1^i = \delta(W_1^i h_0^i), \quad (2)$$

where $W_1^i \in \mathbb{R}^{d_l \times d_h}$ is used to reducing the dimension to $d_h = d_l / r$, and r is the reduction ratio. δ refers to the GeLU activation function.

Weight Generation. To guide the active perception of the visual backbone, we generate language-adaptive weights for producing the query X_q, key X_k, and value X_v in the visual backbone conditioned on referring expressions, which can be represented as:

$$X_q = \theta(X; W_q), X_k = \theta(X; W_k), X_v = \theta(X; W_v), \quad (3)$$

where $\theta(\cdot; W)$ indicates the linear projection operation parameterized by W, and X represents the input visual features. $W_q, W_k, W_v \in \mathbb{R}^{d_{out} \times d_{in}}$ are the dynamic projection weights used to generate the query, key, and value, respectively. d_{in} and d_{out} are the dimension of feature X and query/key/value, respectively.

Considering the large number $d_{out} \times d_{in}$ of the dynamic weights, it is unaffordable to directly generate weights using fully-connected layers like Hypernetworks [10]. The DynamicConv [3] and CondConv [41] can alleviate this problem by generating weights with weighted summation of K static kernels but can increase the parameter number by K-times and suffer from challenging joint optimization. Inspired by the dynamic channel fusion [24], we try to generate dynamic weights following the matrix decomposition paradigm. Taking the i-th ViT block as an example, which can be formulated as:

$$[W_q^i, W_k^i, W_v^i] = W_0^i + P\Phi(h_1^i)Q^T, \quad (4)$$

where $W_0^i \in \mathbb{R}^{d_{out} \times d_{in}}$ is the layer-specific static learnable weights. $P \in \mathbb{R}^{d_{out} \times d_w}$ and $Q \in \mathbb{R}^{d_{in} \times d_w}$ are also static learnable weights, but sharable across all ViT blocks to reduce the parameter numbers and prevent the model from overfitting. $\Phi(h_1^i)$ is a fully-connected layer, which produces a dynamic matrix of shape $d_w \times d_w$ with aggregated linguistic features h_1^i as input.

3.3. Multi-task Head

Different from the previous methods [6, 23, 42, 45, 46, 49, 50], which require carefully designed cross-modal interaction modules, VG-LAW can obtain expression-relevant visual features extracted by the language-aware visual backbone without additional cross-modal interaction modules. Through our proposed neat but efficient multi-task head, we can utilize the visual and linguistic features to predict the bounding box for REC and the segmentation mask for RES. Concretely, there are two branches in the multi-task head for REC and RES, respectively.

For the REC branch, we apply direct coordinate regression to predict the bounding box of referred object. To pool the 2-d visual features along the spatial dimension, we propose a language adaptive pooling module (LAP), which aggregates visual features using language-adaptive attention. Specifically, the visual features $\{F_v^{i,j}\} \in \mathbb{R}^{C \times \frac{H}{s} \times \frac{W}{s}}$ and linguistic feature $F_l^1 \in \mathbb{R}^{d_l}$ are firstly projected to the lower-dimension space $\mathbb{R}^k$, and the attention weights $A \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s}}$ are calculated as dot-product similarity followed by $\mathrm{Softmax}$ normalization. Then, the visual features are aggregated by calculating the weighted sum with attention weights A. Finally, the aggregated visual features are fed to a three-layer fully-connected layer, and the Sigmoid function is used to predict the referred bounding box $\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$.

For the RES branch, we apply binary classification to each visual feature along the spatial dimension to predict segmentation masks for referred objects. Specifically, the visual features F_v are first up-sampled to $\hat{F}_v \in \mathbb{R}^{d_l \times \frac{H}{4} \times \frac{W}{4}}$ with successive transposed convolutions. Then, the intermediate segmentation map $\bar{s} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4}}$ can be obtained by using linear projection $\theta(\cdot; W)$ on each visual feature. Following the language adaptive weight paradigm, we also use dynamic rather than fixed weights by simply setting $W = F_l^1$. Finally, the full-resolution segmentation mask $\hat{s} \in \mathbb{R}^{H \times W}$ is derived by simply up-sample $\bar{s}$ using bilinear interpolation, followed by the Sigmoid function.

3.4. Training Objectives

The VG-LAW framework can be optimized end-to-end for multi-task visual grounding. For REC, given the predicted bounding box $\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$ and the ground truth $b = (x, y, w, h)$, the detection loss function is defined as follows:

$$\mathcal{L}_{det} = \lambda_{L1}\mathcal{L}_{L1}(b, \hat{b}) + \lambda_{giou}\mathcal{L}_{giou}(b, \hat{b}), \tag{5}$$

where $\mathcal{L}_{L1}(\cdot,\cdot)$ and $\mathcal{L}_{giou}(\cdot,\cdot)$ represent L1 loss and Generalized IoU loss [39], respectively, and λ_{L1} and λ_{giou} are the relative weights to control the two detection loss functions. For RES, given the predicted mask $\hat{s}$ and the ground-truth s, the segmentation loss function is defined as follows:

$$\mathcal{L}_{seg} = \lambda_{focal}\mathcal{L}_{focal}(s, \hat{s}) + \lambda_{dice}\mathcal{L}_{dice}(s, \hat{s}), \tag{6}$$

where $\mathcal{L}_{focal}(\cdot,\cdot)$ and $\mathcal{L}_{dice}(\cdot,\cdot)$ represent focal loss [27] and DICE/F-1 loss [35], respectively, and λ_{focal} and λ_{dice} are the relative weights to control the two segmentation loss functions. Our framework can be seamlessly used for joint training of REC and RES, and its joint training loss function is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{seg}. \tag{7}$$

The trained model performs well for language-guided detection and segmentation. The experimental analysis of the whole framework will be elaborated in Sec. 4.

4. Experiments

In this section, we will give a detailed experimental analysis of the whole framework, including the datasets, evaluation protocol, implementation details, comparisons with the state-of-the-art methods, and ablation analysis.

4.1. Datasets and Evaluation Protocol

Datasets. To verify the effectiveness of our method, we conduct experiments on the widely used RefCOCO [47], RefCOCO+ [47], RefCOCOg [34], and ReferItGame [19] datasets. RefCOCO, RefCOCO+, and RefCOCOg are collected from MS-COCO [28]. RefCOCO and RefCOCO+, which are collected in interactive games, can be divided into train, val, testA, and testB sets. Compared to RefCOCO, the expressions of RefCOCO+ contain more attributes than absolute locations. Unlike RefCOCO and RefCOCO+, RefCOCOg collected by Amazon Mechanical Turk has a longer length of 8.4 words, including the attribute and location of referents. Following a common version of split [36], RefCOCOg has train, val, and test sets. In addition, ReferItGame collected from SAIAPR-12 [8] contains train and test sets. Each sample in the above datasets contains its corresponding bounding box and mask.

Evaluation Protocol. Following the previous works [23, 33, 50], we use *Prec@0.5* and *mIoU* to evaluate the performance of REC and RES, respectively. For *Prec@0.5*, the predicted bounding box is considered correct if the intersection-over-union (IoU) with the ground-truth bounding box is greater than 0.5. *mIoU* represents the IoU between the prediction and ground truth averaged across all test samples.

4.2. Implementation Details

Training. The resolution of the input image is resized to 448×448. ViT-Base [7] is used as the visual backbone, and we follow the adaptation introduced by ViTDet [25] to adapt the visual backbone to higher-resolution images. The visual backbone is pre-trained using Mask R-CNN [11] on MS-COCO [28], where overlapping images of the val/test sets are excluded. The W_0^i and $\Phi(h_1^i)$ in Eq. (4) are initialized with the corresponding pre-trained weights of the visual backbone and zeros, respectively. The maximum length of referring expression is set to 40, and the uncased base of six-layer BERT [5] as the linguistic backbone is used to generate linguistic features. λ_{L1} and λ_{giou} are set to 1. λ_{focal} and λ_{dice} are set to 4. The reduction ratio r is set to 16. The initial learning rate for the visual and linguistic backbone is 4e-5, and the initial learning rate for the

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

Table 1. Comparison with state-of-the-art methods on RefCOCO [47], RefCOCO+ [47], RefCOCOg [36] and ReferItGame [19] for REC task. The visual backbone is pre-trained on MS-COCO [28], where overlapping images of the val/test sets are excluded. * represents ImageNet [21] pre-training. 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.

Methods	Venue	Visual Backbone	Multi-task	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg 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

Table 2. Comparison with state-of-the-art methods on RefCOCO [47], RefCOCO+ [47], and RefCOCOg [36] for RES task. The visual backbone is pre-trained on MS-COCO [28], where overlapping images of the val/test sets are excluded. * represents ImageNet [21] 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.

remaining components is 4e-4. The model is end-to-end optimized by AdamW [31] for 90 epochs with a batch size of 256, where weight decay is set to 1e-4, and the learning rate is reduced by a factor of 10 after 60 epochs. Data augmentation operation includes random horizontal flips. We implement our framework using PyTorch and conduct experiments with NVIDIA A100 GPUs.

Inference. At inference time, the input image is resized to 448×448, and the maximum length of referring expressions is set to 40. Following the previous method [33], We set the threshold to 0.35 to realize the binarization of the RES prediction. Without any post-processing operation, our framework directly outputs bounding boxes and segmentation maps specified by referring expressions.

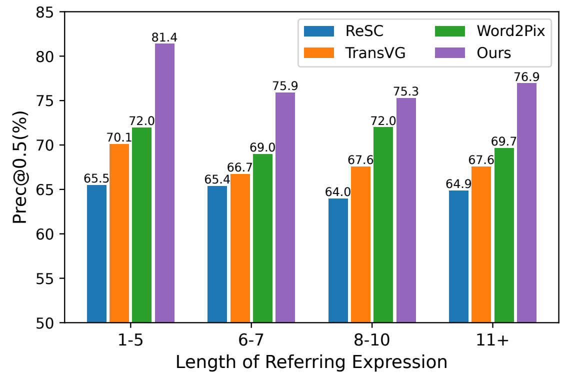

Figure 5. Comparison of accuracy under different lengths of referring expression on RefCOCOg-test. ReSC [42], TransVG [4], Word2Pix [48], and the proposed VG-LAW are compared.

4.3. Comparisons with State-of-the-art Methods

To estimate the effectiveness of the proposed VG-LAW framework, we conduct quantitative experiments on four widely used datasets, *i.e.,* RefCOCO [47], RefCOCO+ [47], RefCOCOg [34], and ReferItGame [19].

REC Task. For the REC task, we compare the performance with state-of-the-art REC methods, including the two-stage methods [2, 13, 30, 46], one-stage methods [26, 33, 42, 43, 49], and transformer-based methods [4, 12, 23, 45, 48, 50]. The main results are summarized to Tab. 1. It can be observed that VG-LAW achieves a significant performance improvement compared to the state-of-the-art two-stage method Ref-NMS [2] and one-stage method PLV-FPN [26]. When comparing to the transformer-based method QRNet [45], which modified the visual backbone by inserting language-aware spatial and channel attention modules, our method has better performance with +2.62%/ +3.47%/ +0.82% on RefCOCO, +3.43%/ +4.87%/ +3.69% on RefCOCO+, +5.01%/ +3.93% on RefCOCOg, and +2.61% on ReferItGame. QRNet [45] follows the TransVG [4] framework, both of which use the transformer encoder-based cross-modal interaction module. Compared to them, VG-LAW achieves better performance without complex cross-modal interaction modules. Furthermore, our method significantly outperforms MCN [33] and RefTR [23] based on joint training of REC and RES.

RES Task. For the RES task, we compare the performance with state-of-the-art methods [6,17,23,32,33,44,50], and the main results are summarized to Tab. 2. Compared with state-of-the-art RES method LAVT [44], VG-LAW achieves better *mIoU* with +1.16%/ +0.62%/ +1.95% on RefCOCO, +2.01%/ +2.42% on RefCOCOg, and comparable *mIoU* with +0.82%/ -0.59%/ -0.34% on RefCOCO+.

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

Table 3. Ablation experiments on ReferItGame [19] to evaluate the proposed language adaptive weight generation (LAWG), language adaptive pooling (LAP), and multi-task head (MTH).

When comparing the models trained with or without multi-task settings, it can also be observed that consistent performance gains are achieved across all the datasets and splits. As REC can provide localization information of the referred object, such coarse-grained supervision can slightly improve the segmentation accuracy in RES.

Analysis of Referring Expression Length. As the visual backbone in VG-LAW extracts features purely perceptually, it is of concern whether it can handle long and complex referring expressions. ReSC [42] reveals that one-stage methods may ignore detailed descriptions in complex referring expressions and lead to poor performance. Following that, we evaluate the REC performance on referring expressions of different lengths, as illustrated in Fig. 5. VG-LAW performs better than ReSC, TransVG [4] and Word2Pix [48], with no significant performance degradation when the length of referring expressions varies from 6-7 to 11+.

4.4. Ablation Analysis

To validate the effectiveness of our proposed modules, *i.e.* language-adaptive weight generation, language-adaptive pooling, and multi-task head, we conduct ablation experiments on the REC dataset of ReferItGame, which is summarized in Tab. 3. When only using the LAWG, the visual features are pooled with global average pooling, and when only using the LAP, the visual backbone has fixed architecture and weights. When only using the LAWG or the LAP, it can be observed that the model already achieves 74.89% and 74.37%, respectively, which is close to the 74.61% reported by QRNet [45]. When combined with the LAWG and LAP, further improvements can be brought by LAWG and LAP with +2.23% and +1.71%, respectively. Benefiting from the auxiliary supervision of RES, our model equipped with the multi-task head can localize the referred objects better and achieve 77.22%.

4.5. Qualitative Results

The qualitative results of the four datasets are shown in Fig. 6. It can be observed that our model can successfully locate and segment the referred objects, and the attention of

Figure 6. Qualitative results on the RefCOCO [47], RefCOCO+ [47], RefCOCOg [34], and ReferItGame [19] datasets. Each dataset shows two examples. 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.

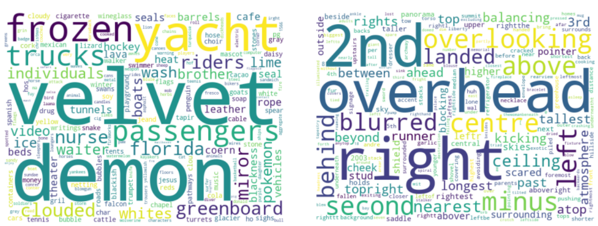

Figure 7. Wordcloud visualization of words assigned to the first and second halves of the visual backbone.

the visual backbone can focus on the most relevant image regions, demonstrating the effectiveness of using language adaptive weights. Taking the results on ReferItGame as an example, the visual backbone can dynamically filter out irrelevant regions for different expressions. For instance, when the "purple curtain" is referred to, the regions related to the "the painting/photo on the wall" are ignored.

In addition, we count the scores of words assigned to the first and second halves of the visual backbone, as shown in Fig. 7. The scores are calculated by averaging attention score α_i in Eq. (1) for each word, followed by softmax normalization along the layer dimension. It can be observed that the shallow layers tend to the words describing individuals, such as the categories "velvet" and "yacht", and the deep layers tend to the words about contexts, such as the ordinal number "2nd" and the position "right".

5. Conclusions and Liminations

In this paper, we propose an active perception framework VG-LAW for visual grounding, based on the language adaptive weights. VG-LAW can directly inject the information of referring expressions into the weights of the visual backbone without modifying its architecture. Equipped with the proposed neat yet efficient multi-task head, VG-LAW achieves state-of-the-art performance for REC and RES tasks on widely used datasets. The limitations of our method are two-fold: (1) VG-LAW is weak in interpretability, and the entire reasoning process is implicit, which makes it difficult to understand how the reasoning process works, and (2) the multi-task head predicts one instance at a time, which limits its application in phrase grounding.

6. Acknowledgments

This work is supported in part by National Key Research and Development Program of China under Grant 2020AAA0107400, National Natural Science Foundation of China under Grant U20A20222, National Science Foundation for Distinguished Young Scholars under Grant 62225605, Ant Group through CCF-Ant Research Fund, and sponsored by CCF-AFSG Research Fund, CAAI-HUAWEI MindSpore Open Fund, CCF-Zhipu AI Large Model Fund(CCF-Zhipu202302) as well as Hikvision Cooperation Fund.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1

[2] Long Chen, Wenbo Ma, Jun Xiao, Hanwang Zhang, and Shih-Fu Chang. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *AAAI*, volume 35, pages 1036–1044, 2021. 6, 7

[3] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, pages 11030–11039, 2020. 3, 4

[4] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, pages 1769–1779, 2021. 1, 2, 6, 7

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 5

[6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *CVPR*, pages 16321–16330, 2021. 1, 3, 4, 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[8] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010. 5

[9] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, pages 15506–15515, 2021. 3

[10] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 3, 4

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1, 5

[12] Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. Yoro-lightweight end to end visual grounding. In *ECCV2022 Workshops*, pages 3–23. Springer, 2023. 2, 6, 7

[13] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE TPAMI*, 2019. 2, 6, 7

[14] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124, 2016. 1, 3

[15] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10488–10497, 2020. 3

[16] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *NeurIPS*, 29, 2016. 3

[17] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021. 1, 3, 6, 7

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 2

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 2, 5, 6, 7, 8

[20] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, pages 18145–18154, 2022. 3

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 6

[22] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*, 2022. 3

[23] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34:19652–19664, 2021. 1, 3, 4, 5, 6, 7

[24] Yunsheng Li, Yinpeng Chen, Xiyang Dai, Dongdong Chen, Ye Yu, Lu Yuan, Zicheng Liu, Mei Chen, Nuno Vasconcelos, et al. Revisiting dynamic convolution via matrix decomposition. In *ICLR*, 2020. 3, 4

[25] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 5

[26] Yue Liao, Aixi Zhang, Zhiyuan Chen, Tianrui Hui, and Si Liu. Progressive language-customized visual feature learning for one-stage visual grounding. *IEEE TIP*, 31:4266–4277, 2022. 6, 7

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 5

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 5, 6

[29] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, pages 4673–4682, 2019. 2

[30] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, pages 1950–1959, 2019. 2, 6, 7

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *ACM MM*, pages 1274–1282, 2020. 1, 3, 6, 7

[33] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 1, 5, 6, 7

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 5, 7, 8

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision*, pages 565–571. IEEE, 2016. 5

[36] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 2, 5, 6

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 2

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2, 3, 4

[41] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *NeurIPS*, 32, 2019. 3, 4

[42] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *ECCV*, pages 387–404, 2020. 1, 2, 4, 6, 7

[43] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, pages 4683–4693, 2019. 2, 6, 7

[44] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 1, 2, 3, 6, 7

[45] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *CVPR*, pages 15502–15512, 2022. 1, 2, 4, 6, 7

[46] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 1, 2, 4, 6, 7

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 2, 5, 6, 7, 8

[48] Heng Zhao, Joey Tianyi Zhou, and Yew-Soon Ong. Word2pix: Word to pixel cross-attention transformer in visual grounding. *IEEE Trans. Neural Networks and Learning Systems.*, 2022. 1, 2, 6, 7

[49] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Trans. Neural Networks and Learning Systems.*, 2021. 2, 4, 6, 7

[50] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. *arXiv preprint arXiv:2203.16265*, 2022. 1, 2, 3, 4, 5, 6, 7