# GRES: Generalized Referring Expression Segmentation

Chang Liu[†]     Henghui Ding[† ✉]     Xudong Jiang
Nanyang Technological University, Singapore
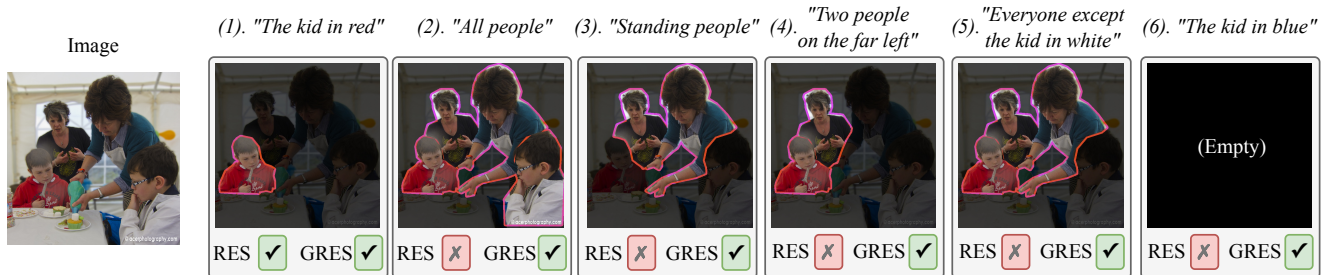https://henghuiding.github.io/GRES

Figure 1. Classic Referring Expression Segmentation (RES) only supports expressions that indicate a single target object, *e.g.*, (1). Compared with classic RES, the proposed **Generalized Referring Expression Segmentation (GRES)** supports expressions indicating an *arbitrary number* of target objects, for example, multi-target expressions like (2)-(5), and no-target expressions like (6).

## Abstract

*Referring Expression Segmentation (RES) aims to generate a segmentation mask for the object described by a given language expression. Existing classic RES datasets and methods commonly support single-target expressions only, i.e., one expression refers to one target object. Multi-target and no-target expressions are not considered. This limits the usage of RES in practice. In this paper, we introduce a new benchmark called Generalized Referring Expression Segmentation (GRES), which extends the classic RES to allow expressions to refer to an arbitrary number of target objects. Towards this, we construct the first large-scale GRES dataset called gRefCOCO that contains multi-target, no-target, and single-target expressions. GRES and gRefCOCO are designed to be well-compatible with RES, facilitating extensive experiments to study the performance gap of the existing RES methods on the GRES task. In the experimental study, we find that one of the big challenges of GRES is complex relationship modeling. Based on this, we propose a region-based GRES baseline ReLA that adaptively divides the image into regions with sub-instance clues, and explicitly models the region-region and region-language dependencies. The proposed approach ReLA achieves new state-of-the-art performance on the both newly proposed GRES and classic RES tasks. The proposed gRefCOCO dataset and method are available at https://henghuiding.github.io/GRES.*

## 1. Introduction

Referring Expression segmentation (RES) is one of the most important tasks of multi-modal information processing. Given an image and a natural language expression that describes an object in the image, RES aims to find this target object and generate a segmentation mask for it. It has great potential in many applications, such as video production, human-machine interaction, and robotics. Currently, most of the existing methods follow the RES rules defined in the popular datasets ReferIt [20] and RefCOCO [34, 47] and have achieved great progress in recent years.

**Limitations of classic RES.** However, most classic RES methods have some strong pre-defined constraints to the task. First, the classic RES does not consider no-target expressions that do not match any object in the image. This means that the behavior of the existing RES methods is undefined if the target does not exist in the input image. When it comes to practical applications under such constraint, the input expression has to match an object in the image, otherwise problems inevitably occur. Second, most existing datasets, *e.g.*, the most popular RefCOCO [34, 47], do not contain multi-target expressions that point to multiple instances. This means that multiple inputs are needed to search objects one by one. *E.g.*, in Fig. 1, four distinct expressions with four times of model calls are

---

[†]Equal contribution.
✉Corresponding author (henghui.ding@gmail.com).

Table 1. Comparison among different referring expression datasets, including ReferIt [20], RefCOCO(g) [34, 47], PhraseCut [40], and our proposed **gRefCOCO**. Multi-target: expression that specifies multiple objects in the image. No-target: expression that does not touch on any object in the image.

|  | ReferIt | RefCOCO(g) | PhraseCut | **gRefCOCO** |
|---|---|---|---|---|
| Image Source | CLEF [8] | COCO [25] | VG [22] | COCO [25] |
| Multi-target | ✗ | ✗ | (fallback) | ✓ |
| No-target | ✗ | ✗ | ✗ | ✓ |
| Expression type | free | free | templated | free |

required to segment *"All people"*. Our experiments show that classic RES methods trained on existing datasets cannot be well-generalized to these scenarios.

**New benchmark and dataset.** In this paper, we propose a new benchmark, called Generalized Referring Expression Segmentation (GRES), which allows expressions indicating any number of target objects. GRES takes an image and a referring expression as input, the same as classic RES. Different from classic RES, as shown in Fig. 1, GRES further supports multi-target expression that specifies multiple target objects in a single expression, *e.g.*, *"Everyone except the kid in white"*, and no-target expression that does not touch on any object in the image, *e.g.*, *"the kid in blue"*. This provides much more flexibility for input expression, making referring expression segmentation more useful and robust in practice. However, existing referring expression datasets [20, 34, 47] do not contain multi-target expression nor no-target samples, but only have single-target expression samples, as shown in Tab. 1. To facilitate research efforts on realistic referring segmentation, we build a new dataset for GRES, called gRefCOCO. It complements RefCOCO with two kinds of samples: multi-target samples, in which the expression points to two or more target instances in the image, and no-target samples, in which the expression does not match any object in the image.

**A baseline method.** Moreover, we design a baseline method based on the objectives of the GRES task. It is known that modeling relationships, *e.g.*, region-region interactions, plays a crucial role in RES [46]. However, classic RES methods only have one target to detect so that many methods can achieve good performance without explicit region-region interaction modeling. But in GRES, as multi-target expressions involve multiple objects in one expression, it is more challenging and essential to model the long-range region-region dependencies. From this point, we propose a region-based method for GRES that explicitly model the interaction among regions with sub-instance clues. We design a network that splits the image into regions and makes them explicitly interact with each other. Moreover, unlike previous works where regions come from a simple hard-split of the input image, our network soft-collates features for each region, achieving more flexibility. We do extensive experiments on our

proposed methods against other RES methods, showing that the explicit modeling of interaction and flexible region features greatly contributes to the performance of GRES.

In summary, our contributions are listed as follows:

1. We propose a benchmark of Generalized Referring Expression Segmentation (GRES), making RES more flexible and practical in real-world scenarios.

2. We propose a large-scale GRES dataset gRefCOCO. To the best of our knowledge, this is the first referring expression dataset that supports expressions indicating an arbitrary number of target objects.

3. We propose a solid baseline method ReLA for GRES to model complex **ReLA**tionships among objects, which achieves the new state-of-the-art performance on both classic RES and newly proposed GRES tasks.

4. We do extensive experiments and comparisons of the proposed baseline method and other existing RES methods on the GRES, and analyze the possible causes of the performance gap and new challenges in GRES.

## 2. Related Works

**Related referring tasks and datasets.** Being defined by Hu *et al*. [13], Referring Expression *Segmentation* (RES) comes from a similar task, Referring Expression *Comprehension* (REC) [14, 24, 27, 38, 42, 43, 52] that outputs a bounding box for the target. The earliest dataset for RES and REC is ReferIt [20], in which one expression only refers to one instance. Later, Yu *et al*. propose RefCOCO [47] for RES and REC. However, like ReferIt, it only contains single-target expressions. Another popular dataset RefCOCOg [34] also inherits this. Although the original definition of RES [13] does not limit the number of target instances, "**one expression, one instance**" has become a "de-facto" rule for RES task.

Recently, some new datasets are proposed, but most of them are neither focused on nor suitable for GRES. *E.g.*, although PhraseCut [40] has multi-target expressions, it only considers them as "fallback", *i.e.*, multi-target expressions are only used when an object cannot be uniquely referred to. In contrast, our expression intentionally finds multiple targets. Besides, expressions in PhraseCut are written using templates rather than free natural language expressions, limiting the diversity of language usage. Image caption datasets [22, 36] are close to RES, but they cannot ensure unambiguity of expression→object(s). Thus, they are not suitable for referring-related tasks. There are some referring datasets using other data modalities or learning schemes, *e.g.*, Scanrefer [2] focuses on 3D objects and Clevrtex [19] focuses on unsupervised learning. Moreover, none of the above datasets has no-target expression.

(a) Multi-target: Selecting multiple objects in one single forward process.

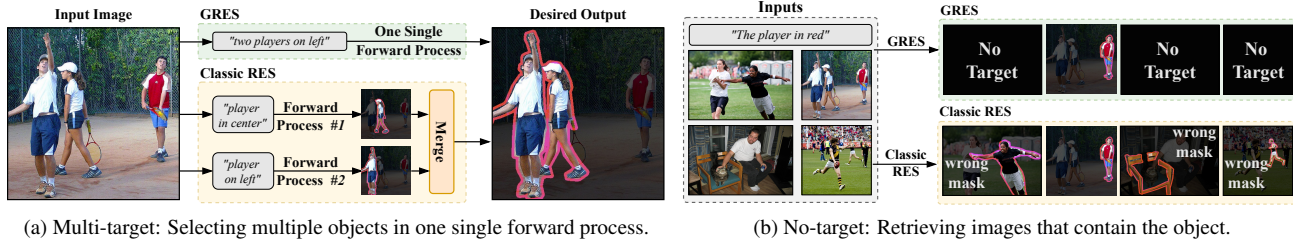(b) No-target: Retrieving images that contain the object.

Figure 2. More applications of GRES brought by supporting multi-target and no-target expressions compared to classic RES.

**Referring segmentation methods.** Referring segmentation methods can be roughly divided into two categories: one-stage (or top-down) methods [1, 4, 15–17, 23, 31, 35, 45, 51] and two-stage (or bottom-up) methods [26, 46]. One-stage methods usually have an FCN-like [30] end-to-end network, and the prediction is achieved by per-pixel classification on fused multi-modal feature. Two-stage methods first find a set of instance proposals using an out-of-box instance segmentation network and then select the target instance from them. The majority of RES methods are one-stage, while two-stage methods are more prevalent in REC [12, 14, 28, 33, 48, 50]. Most recently, some transformer-based methods [5, 6, 39, 44] are proposed and bring large performance gain compared to the CNN-based network. Zero-shot segmentation methods [9, 10, 49] use class names as textual information and focus on identifying novel categories, in contrast to RES that employs natural expressions to identify the user's intended target.

## 3. Task Setting and Dataset

### 3.1. GRES Settings

**Revisit of RES.** Classic Referring Expression Segmentation (RES) takes an image and an expression as inputs. The desired output is a segmentation mask of the target region that is referred by the input expression. As discussed in Sec. 2, the current RES does not consider no-target expressions, and all samples in current datasets only have single-target expressions. Thus, existing models are likely to output an instance incorrectly if the input expression refers to nothing or multiple targets in the input image.

**Generalized RES.** To address these limitations in classic RES, we propose a benchmark called Generalized Referring Expression Segmentation (GRES) that allows expressions indicating arbitrary number of target objects. A GRES data sample contains four items: an image $I$, a language expression $T$, a ground-truth segmentation mask $M_{GT}$ that covers pixels of all targets referred by $T$, and a binary no-target label $E_{GT}$ that indicates whether $T$ is a no-target expression. The number of instances in $T$ is not limited. GRES models take $I$ and $T$ as inputs and predict a mask $M$. For no-target expressions, $M$ should be all negative.

The applications of multi-target and no-target expres-

sions are not only finding multiple targets and rejecting inappropriate expressions matching nothing, but also bringing referring segmentation into more realistic scenarios with advanced usages. For example, with the support of multi-target expressions, we can use expressions like "*all people*" and "*two players on left*" as input to select multiple objects in a single forward process (see Fig. 2a), or use expressions like "*foreground*" and "*kids*" to achieve user-defined open vocabulary segmentation. With the support of no-target expressions, users can apply the same expression on a set of images to identify which images contain the object(s) in the language expression, as in Fig. 2b. This is useful if users want to find and matte something in a group of images, similar to image retrieval but more specific and flexible. What's more, allowing multi-target and no-target expressions enhances the model's reliability and robustness to realistic scenarios where any type of expression can occur unexpectedly, for example, users may accidentally or intentionally mistype a sentence.

**Evaluation.** To encourage the diversity of GRES methods, we do not force GRES methods to differentiate different instances in the expression though our dataset gRefCOCO provides, enabling popular one-stage methods to participate in GRES. Besides the regular RES performance metric cumulative IoU (cIoU) and Precision@X, we further propose a new metric called generalized IoU (gIoU), which extends the mean IoU to all samples including no-target ones. Moreover, No-target performance is also separately evaluated by computing No-target-accuracy (N-acc.) and Target-accuracy (T-acc.). Details are given in Sec. 5.1.

### 3.2. gRefCOCO: A Large-scale GRES Dataset

To perform the GRES task, we construct the gRefCOCO dataset. It contains 278,232 expressions, including 80,022 multi-target and 32,202 no-target expressions, referring to 60,287 distinct instances in 19,994 images. Masks and bounding boxes for all target instances are given. Part of single-target expressions are inherited from RefCOCO. We developed an online annotation tool to find images, select instances, write expressions, and verify the results.

The basic annotation procedure follows ReferIt [20] to ensure the annotation quality. The data split is also kept the same as the UNC partition of RefCOCO [47]. We compare

| Image (a) | i. *"The **two** people on the far left"* | ii. *"Everyone **except** the kid in white"* | Image (b) | i. *"The bike **and** two passengers on **it**"* | ii. *"The bike **that has** two passengers and **its** driver"* |

Figure 3. Examples of the proposed gRefCOCO dataset.

the proposed gRefCOCO with RefCOCO and list some unique and significant features of our dataset as follows.

**Multi-target samples.** In practice, users usually cluster multiple targets of an image by describing their logical relationships or similarities. From this point, we let annotators select target instances rather than randomly assembling them. Then annotators write an unambiguous referring expression for the selected instances. There are four major features and challenges brought by multi-target samples:

**1) Usage of counting expressions**, *e.g.*, "*The **two** people on the far left*" in Fig. 3(a). As the original RefCOCO already has ordinal word numbers like "*the second person from left*", the model must be able to differentiate cardinal numbers from ordinal numbers. Explicit or implicit object-counting ability is desired to address such expressions.

**2) Compound sentence structures without geometrical relation**, like compound sentences "*A and B*", "*A except B*", and "*A with B or C*", as shown in Fig. 3. This raises higher requirements for models to understand the long-range dependencies of both the image and the sentence.

**3) Domain of attributes.** When there are multiple targets in an expression, different targets may share attributes or have different attributes, *e.g.*, "*the right lady in blue and kid in white*". Some attributes may be shared, *e.g.*, "*right*", and others may not, *e.g.*, "*blue*" and "*white*". This requires the model to have a deeper understanding of all the attributes and map the relationship of these attributes to their corresponding objects.

**4) More complex relationships**. Since a multi-target expression involves more than one target, relationship descriptions appear more frequently and are more complicated than in sing-target ones. Fig. 3(b) gives an example. Two similar expressions are applied on the same image. Both expressions have the conjunction word "*and*", and "*two passengers*" as an attribute to the target "*bike*". But the two expressions refer to two different sets of targets as shown in Fig. 3(b). Thus in GRES, relationships are not only used to describe the target but also indicate the number of targets. This requires the model to have a deep understanding of all instances and their interactions in the image and expression.

**No-target samples.** During the annotation, we found that if we do not set any constraints for no-target expressions, annotators tend to write a lot of simple or general expressions that are quite different from other expressions with valid targets. *E.g.*, annotators may write duplicated "*dog*" for all images without dogs. To avoid these undesirable and purposeless samples in the dataset, we set two rules for no-target expressions:

**1) The expression cannot be totally irrelevant to the image**. For example, given the image in Fig. 3(a), "*The kid in blue*" is acceptable as there do exist kids in the image, but none of them is in blue. But expressions like "*dog*", "*car*", "*river*" *etc*. are unacceptable as they are totally irrelevant to anything in this image.

**2) The annotators could choose a deceptive expression** drawn from other images in RefCOCO's same data split, if an expression required by in 1) is hard to come up with.

These rules greatly improve the diversity of no-target expressions and keep our dataset at a reasonable difficulty. More examples are shown in the Supplementary Materials.

## 4. The Proposed Method for GRES

As discussed earlier, the relationship and attribute descriptions are more complex in multi-target expressions. Compared with classic RES, it is more challenging and important for GRES to model the complex interaction among regions in the image, and capture fine-grained attributes for all objects. We propose to explicitly interact different parts of image and different words in expression to analyze their dependencies.

### 4.1. Architecture Overview

The overview of our framework is shown in Fig. 4. The input image is processed by a transformer encoder based on Swin [29] to extract vision features $F_i \in \mathbb{R}^{H \times W \times C}$, in which $H, W$ are the spatial size and $C$ is the channel dimensions. The input language expression is processed by BERT [3], producing the language feature $F_t \in \mathbb{R}^{N_t \times C}$, where $N_t$ is the number of words in the expression. Next, $F_i$ is sent to a pixel decoder to obtain the mask feature $F_m$ for mask prediction. Meanwhile, $F_i$ and $F_t$ are sent to our proposed **ReLA**tionship modeling block (see Sec. 4.2 for details), which divides the feature maps into $P \times P = P^2$ regions, and models the interaction among them. These "regions" correspond to the image's $P \times P$ patches like ViT [7]. However, the shape and sizes of their spatial areas are
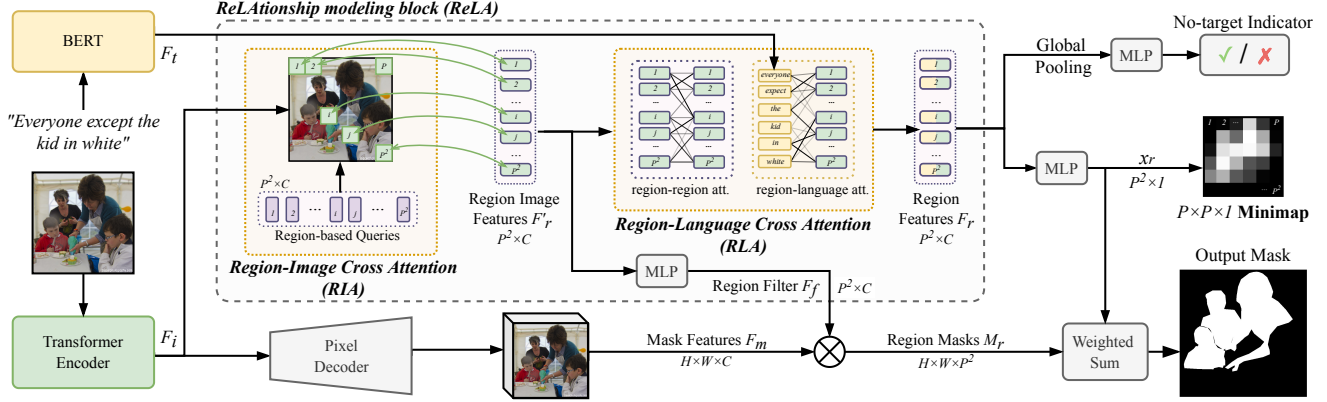
Figure 4. Architecture overview of the GRES baseline model **ReLA**. Firstly, the given image and expression are encoded into vision feature $F_i$ and language feature $F_t$, respectively. $F_i$ is fed into a pixel decoder to produce mask features $F_m$. **ReLA**tionship modeling block takes both $F_i$ and $F_t$ as inputs and output 1) region filter $F_f$ that produces region masks $M_r$, 2) region probability map $x_r$, and 3) no-target judgement score $E$. Output mask is obtained by weighted fusion of region masks $M_r$.

not predefined but found by ReLA dynamically, which is different from previous works using hard-split [7,21,37,41]. ReLA generates two sets of features: region feature $F_r = \{f_r^n\}_{n=1}^{P^2}$ and region filter $F_f = \{f_f^n\}_{n=1}^{P^2}$. For the $n$-th region, its region feature $f_r^n$ is used to find a scalar $x_r^n$ that indicates its probability of containing targets, and its region filter $f_f^n$ is multiplied with the mask feature $F_m$, generating its regional segmentation mask $M_r^n \in \mathbb{R}^{H \times W}$ that indicates the area of this region. We get the predicted mask by weighted aggregating these masks:

$$M = \sum_n (x_r^n M_r^n). \tag{1}$$

**Outputs and Loss.** The predicted mask $M$ is supervised by the ground-truth target mask $M_{GT}$. The $P \times P$ probability map $x_r$ is supervised by a "minimap" downsampled from $M_{GT}$, so that we can link each region with its corresponding patch in the image. Meantime, we take the global average of all region features $F_r$ to predict a no-target label $E$. In inference, if $E$ is predicted to be positive, the output mask $M$ will be set to empty. $M$, $x_r$ and $E$ are guided by the cross-entropy loss.

### 4.2. ReLAtionship Modeling

The proposed **ReLA**tionship modeling has two main modules, Region-Image Cross Attention (RIA) and Region-Language Cross Attention (RLA). The RIA flexibly collects region image features. The RLA captures the region-region and region-language dependency relationships.
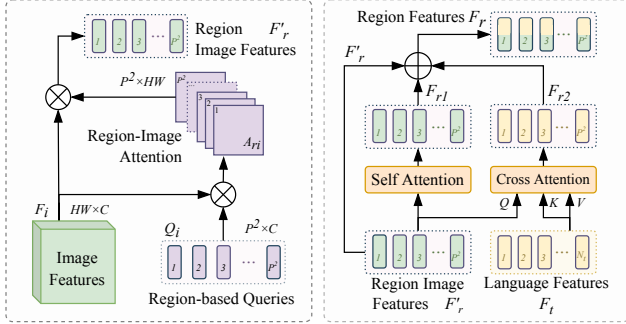
**Region-Image Cross Attention (RIA).** RIA takes the vision feature $F_i$ and $P^2$ learnable Region-based Queries $Q_r$ as input. Supervised by the minimap shown in Fig. 4, each query corresponds to a spatial region in the image and is responsible for feature decoding of the region. The architecture is shown in Fig. 5a. First, the attention between image feature $F_i$ and $P^2$ query embeddings $Q_r \in \mathbb{R}^{P^2 \times C}$

is performed to generate $P^2$ attention maps:

$$A_{ri} = \text{softmax}(Q_r \sigma(F_i W_{ik})^T), \tag{2}$$

where $W_{ik}$ is $C \times C$ learnable parameters and $\sigma$ is GeLU [11]. The resulting $A_{ri} \in \mathbb{R}^{P^2 \times HW}$ gives each query a $H \times W$ attention map indicating its corresponding spatial areas in the image. Next, we get the region features from their corresponding areas using these attention maps: $F_r' = A_{ri}\sigma(F_i W_{iv})^T$, where $W_{iv}$ is $C \times C$ learnable parameters. In such a way, the feature of each region is dynamically collected from their relevant positions. Compared to hard-splitting the image into patches, this method gives more flexibility. An instance may be represented by multiple regions in the minimap (see Fig. 4), making regions represent more fine-grained attributes at the sub-instance level, *e.g.*, the head and upper body of a person. Such sub-instance representations are desired for addressing the complex relationship and attribute descriptions in GRES. A region filter $F_f$ containing region clues is obtained based on $F_r'$ for mask prediction. $F_r'$ is further fed into RLA for region-region and region-word interaction modeling.

**Region-Language Cross Attention (RLA).** Region image features $F_r'$ come from collating image features that do not contain relationship between regions and language information. We propose RLA module to model the region-region and region-language interactions. As in Fig. 5b, RLA consists of a self-attention for region image features $F_r'$ and a multi-modal cross attention. The self-attention models the region-region dependency relationships. It computes the attention matrix by interacting one region feature with all other regions and outputs the relationship-aware region feature $F_{r1}$. Meanwhile, the cross attention takes language feature $F_t$ as Value and Key input, and region image feature $F_r'$ as Query input. This firstly models the relationship between each word and each region:

(a) RIA                                    (b) RLA

Figure 5. Architectures of Region-Image Cross Attention (RIA) and Region-Language Cross Attention (RLA).

$$A_l = \text{softmax}(\sigma(F_r' W_{lq})\sigma(F_t W_{lk})^T), \qquad (3)$$

where $A_l \in \mathbb{R}^{P^2 \times N_t}$. Then it forms the language-aware region features using the derived word-region attention: $F_{r2} = A_l F_t$. Finally, the interaction-aware region feature $F_{r1}$, language-aware region feature $F_{r2}$, and region image features $F_r'$ are added together, and a MLP further fuses the three sets of features: $F_r = \text{MLP}(F_r' + F_{r1} + F_{r2})$.

# 5. Experiments and Discussion

## 5.1. Evaluation Metrics

Besides the widely-used RES metrics cumulative IoU (cIoU) and Precision@X (Pr@X), we further introduce No-target accuracy (N-acc.), Target accuracy (T-acc.), and generalized IoU (gIoU) for GRES.

**cIoU and Pr@X**. cIoU calculates the total intersection pixels over total union pixels, and Pr@X counts the percentage of samples with IoU higher than the threshold $X$. Notably, no-target samples are excluded in Pr@X. And as multi-target samples have larger foreground areas, models are easier to get higher cIoU scores. Thus, we raise the starting threshold to 0.7 for Pr@X.

**N-acc. and T-acc.** evaluates the model's performance on no-target identification. For a no-target sample, prediction without any foreground pixels is true positive ($TP$), otherwise false negative ($FN$). Then, N-acc. measures the model's performance on identifying no-target samples: N-acc. $= \frac{TP}{TP+FN}$. T-acc. reflects how much the generalization on no-target affects the performance on target samples, *i.e.* how many samples that have targets are misclassified as no-target: T-acc. $= \frac{TN}{TN+FP}$.

**gIoU.** It is known that cIoU favors larger objects [40,44]. As multi-target samples have larger foreground areas in GRES, we introduce generalized IoU (gIoU) that treats all samples equally. Like mean IoU, gIoU calculates the mean value of per-image IoU over all samples. For no-target samples, the IoU values of true positive no-target samples are regarded as 1, while IoU values of false negative samples are treated as 0.



Figure 6. Example predictions of the same model being trained on RefCOCO *vs.* gRefCOCO.

Table 2. Ablation study of RIA design options.

| # | Methods | P@0.7 | P@0.8 | P@0.9 | cIoU | gIoU |
|---|---------|-------|-------|-------|------|------|
| #1 | Hard split, input | 63.02 | 59.81 | 19.26 | 54.43 | 55.34 |
| #2 | Hard split, decoder | 70.34 | 65.23 | 21.47 | 60.08 | 60.93 |
| #3 | w/o minimap | 72.19 | 66.02 | 21.07 | 61.30 | 62.06 |
| #4 | **ReLA** (ours) | **74.20** | **68.33** | **24.68** | **62.42** | **63.60** |

## 5.2. Ablation Study

**Dataset necessity.** To show the necessity and validity of gRefCOCO on the task of GRES, we compare the results of the same model trained on RefCOCO and gRefCOCO. As shown in Fig. 6, image (a) is a multi-target sample using a shared attribute (*"in black jacket"*) to find *"two guys"*. The model trained on RefCOCO only finds one, even though the expression explicitly points out that there are two target objects. Image (b) gives a no-target expression, and the RefCOCO-trained model outputs a meaningless mask. The results demonstrate that models trained only on single-target referring expression datasets, *e.g.*, RefCOCO, cannot be well generalized to GRES. In contrast, the newly built gRefCOCO can effectively enable the model to handle expressions indicating an arbitrary number of objects.
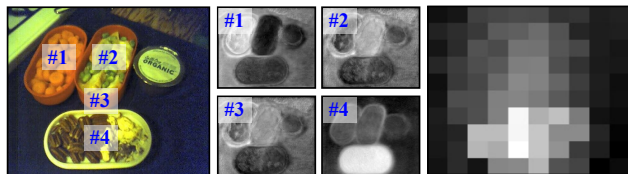
**Design options of RIA.** In Tab. 2, we investigate the performance gain brought by RIA. In model #1, we follow previous methods [7, 21] and rigidly split the image into $P \times P$ patches before sending them into the encoder. Tab. 2 shows that this method is not suitable for our ReLA framework, because it makes the global image information less pronounced due to compromised integrity. In model #2, RIA is replaced by average pooling the image feature into $P \times P$. The gIoU gets a significant gain of $5.59\%$ from model #1, showing the importance of global context in visual feature encoding. Then, another $2.67\%$ gIoU gain can be got by adding our proposed dynamic region feature aggregation for each query (Eq. (2)), showing the effectiveness of the proposed adaptive region assigning. Moreover, we study the importance of linking queries with actual image regions. In model #3, we removed the minimap supervision so that the region-based queries $Q_r$ become plain learnable queries,

Table 3. Ablation study of RLA design options.

| # | Methods | P@0.7 | P@0.8 | P@0.9 | cIoU | gIoU |
|---|---------|-------|-------|-------|------|------|
| #1 | Baseline | 69.94 | 61.10 | 19.38 | 57.24 | 58.53 |
| #2 | + language att. | 72.03 | 65.42 | 21.04 | 59.86 | 60.53 |
| #3 | + region att. | 73.52 | 67.01 | 23.43 | 61.00 | 62.38 |
| #4 | **ReLA** (ours) | **74.20** | **68.33** | **24.68** | **62.42** | **63.60** |

Table 4. Ablation study of Number of Regions

| # Regions | P@0.7 | P@0.8 | P@0.9 | cIoU | gIoU |
|-----------|-------|-------|-------|------|------|
| $4 \times 4$ | 68.48 | 60.25 | 20.33 | 56.57 | 57.01 |
| $8 \times 8$ | 72.36 | 66.85 | 23.56 | 59.74 | 61.23 |
| $10 \times 10$ | **74.20** | **68.33** | **24.68** | **62.42** | **63.60** |
| $12 \times 12$ | 74.14 | 67.56 | 23.90 | 62.02 | 63.50 |



"All three lunch boxes"          Predicted Minimap

Figure 7. Visualization of the predicted minimap & region masks.

resulting in a $1.54\%$ gIoU drop. This shows that explicit correspondence between queries and spatial image regions is beneficial to our network.

**Design options of RLA.** Tab. 3 shows the importance of dependency modeling to GRES. In the baseline model, RLA is replaced by point-wise multiplying region features and globally averaged language features, to achieve a basic feature fusion like previous works [5, 32]. In model #2, the language cross attention is added onto the baseline model, which brings a gIoU gain of $2\%$. This shows the validity of region-word interaction modeling. Then we further add the region self-attention to investigate the importance of the region-region relationship. The region-region relationship modeling brings a performance gain of $3.85\%$ gIoU. The region-region and region-word relationship modeling together bring a significant improvement of $5.07\%$ gIoU.

**Number of regions** $P$. Smaller $P$ leads to coarser regions, which is not good for capturing fine-grained attributes, while larger $P$ costs more resources and decreases the area of each region, making relationship learning difficult. We do experiments on the selection of $P$ in Tab. 4 to find the optimized $P$. The model's performance improves as $P$ increases until 10, which is selected as our setting. In Fig. 7, we visualize the predicted minimap $x_r$ and region maps $M_r$. $x_r$ displays a rough target probability of each region, showing the effectiveness of minimap supervision. We also see that the region masks capture the spatial correlation of the corresponding regions. With flexible region size and shape, each region mask contains not only the instance of this region but also other instances with strong relationships. For example, region #4 is located inside the bottom lunch box, but as the input expression tells that all three boxes are targets, the top two also cause some responses in the output mask of region #4.

Table 5. Comparison on gRefCOCO dataset.

| Methods | val | | testA | | testB | |
|---------|-----|-----|-------|-----|-------|-----|
| | cIoU | gIoU | cIoU | gIoU | cIoU | gIoU |
| MattNet [46] | 47.51 | 48.24 | 58.66 | 59.30 | 45.33 | 46.14 |
| LTS [18] | 52.30 | 52.70 | 61.87 | 62.64 | 49.96 | 50.42 |
| VLT [5] | 52.51 | 52.00 | 62.19 | 63.20 | 50.52 | 50.88 |
| CRIS [39] | 55.34 | 56.27 | 63.82 | 63.42 | 51.04 | 51.79 |
| LAVT [44] | 57.64 | 58.40 | 65.32 | 65.90 | 55.04 | 55.83 |
| VLT+ReLA | 58.65 | 59.43 | 66.60 | 65.35 | 56.22 | 57.36 |
| LAVT+ReLA | 61.23 | 61.32 | 67.54 | 66.40 | 58.24 | 59.83 |
| **ReLA** (ours) | **62.42** | **63.60** | **69.26** | **70.03** | **59.88** | **61.02** |

Table 6. No-target results comparison on gRefCOCO dataset.

| Methods | val | | testA | | testB | |
|---------|-----|-----|-------|-----|-------|-----|
| | N-acc. | T-acc. | N-acc. | T-acc. | N-acc. | T-acc. |
| MattNet [46] | 41.15 | 96.13 | 44.04 | 97.56 | 41.32 | 95.32 |
| VLT [5] | 47.17 | 95.72 | 48.74 | 95.86 | 47.82 | 94.66 |
| LAVT [44] | 49.32 | 96.18 | 49.25 | 95.08 | 48.46 | 95.34 |
| **ReLA**-50pix | 49.96 | 96.28 | 51.36 | 96.35 | 49.24 | 95.02 |
| **ReLA** | **56.37** | **96.32** | **59.02** | **97.68** | **58.40** | **95.44** |

### 5.3. Results on GRES

**Comparison with state-of-the-art RES methods.** In Tab. 5, we report the results of classic RES methods on gRefCOCO. We re-implement these methods using the same backbone as our model and train them on gRefCOCO. For one-stage networks, output masks with less than 50 positive pixels are cleared to all-negative, for better no-target identification. For the two-stage network MattNet [46], we let the model predict a binary label for each instance that indicates whether this candidate is a target, then merge all target instances. As shown in Tab. 5, these classic RES methods do not perform well on gRefCOCO that contains multi-target and no-target samples. Furthermore, to better verify the effectiveness of explicit modeling, we add our ReLA on VLT [5] and LAVT [44] to replace the decoder part of them. From Tab. 5, our explicit relationship modeling greatly enhances model's performance. *E.g.*, adding ReLA improves the cIoU performance of the LAVT by more than $4\%$ on the val set.

In Tab. 6, we test the no-target identification performance. As shown in the table, T-acc. of all methods are mostly higher than $95\%$, showing that our gRefCOCO does not significantly affect the model's targeting performance while being generalized to no-target samples. But from N-acc. of classic RES methods, we see that even being trained with no-target samples, it is not satisfactory to identify no-target samples solely based on the output mask. We also tested our model with the no-target classifier disabled and only use the positive pixel count in the output mask to identify no-target samples ("ReLA-50pix" in Tab. 6). The performance is similar to other methods. This shows that a dedicated no-target classifier is desired. However, although our N-acc. is higher than RES methods, there are still around $40\%$ of no-target samples are missed. We speculate that this is because many no-target expressions are very

Table 7. Results on classic RES in terms of cIoU. U: UMD split. G: Google split.

| Methods | Visual Encoder | Textual Encoder | RefCOCO | | | RefCOCO+ | | | G-Ref | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | val | test A | test B | val | test A | test B | val(U) | test(U) | val(G) |
| MCN [32] | Darknet53 | bi-GRU | 62.44 | 64.20 | 59.71 | 50.62 | 54.99 | 44.69 | 49.22 | 49.40 | - |
| VLT [5] | Darknet53 | bi-GRU | 67.52 | 70.47 | 65.24 | 56.30 | 60.98 | 50.08 | 54.96 | 57.73 | 52.02 |
| ReSTR [21] | ViT-B | Transformer | 67.22 | 69.30 | 64.45 | 55.78 | 60.44 | 48.27 | - | - | 54.48 |
| CRIS [39] | CLIP-R101 | CLIP | 70.47 | 73.18 | 66.10 | 62.27 | 68.08 | 53.68 | 59.87 | 60.36 | - |
| LAVT [44] | Swin-B | BERT | 72.73 | 75.82 | 68.79 | 62.14 | 68.38 | 55.10 | 61.24 | 62.09 | 60.50 |
| VLT [6] | Swin-B | BERT | 72.96 | 75.96 | 69.60 | 63.53 | 68.43 | 56.92 | 63.49 | **66.22** | **62.80** |
| **ReLA** (ours) | Swin-B | BERT | **73.82** | **76.48** | **70.18** | **66.04** | **71.02** | **57.65** | **65.00** | 65.97 | 62.70 |



Figure 8. Example results of our method on gRefCOCO dataset.



Figure 9. Failure cases of our method on gRefCOCO dataset.

deceptive and similar with real instances in the image. We believe that no-target identification will be one of our key focus on the future research for the GRES task.

**Qualitative results.** Some qualitative examples of our model on the val set of gRefCOCO are shown in Fig. 8. In Image (a), our model can detect and precisely segment multiple targets of the same category (*"girls"*) or different categories (*"girls and the dog"*), showing the strong generalization ability. Image (b) uses counting words (*"two bowls"*) and shared attributes (*"on right"*) to describe a set of targets. Image (c) has a compound sentence showing that our model can understand the excluding relationship: *"except the blurry guy"* and makes a good prediction.

**Failure cases & discussion.** We show some failure cases of our method in Fig. 9. Image (a) introduces a possession relationship: *"left girl and **her** laptop"*. This is a very deceptive case. In the image, the laptop in center is more dominant and closer to the left girl than the left one, so the model highlighted the center laptop as *"her laptop"*. Such a challenging case requires the model to have a profound understanding of all objects, and a contextual comprehension of the image and expression. In the second case, the expression is a no-target expression, referring to *"man in gray shirt sitting on bed"*. In the image, there is indeed a sitting person in grey shirt, but he is sitting on a black chair very close to the bed. This further requires the model to look into the fine-grained details of all objects, and understand those details with image context.
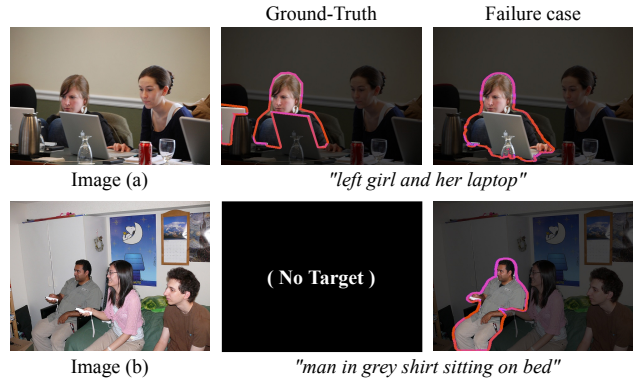
## 5.4. Results on Classic RES

We also evaluate our method on the classic RES task and report the results in Tab. 7. In this experiment, our model strictly follows the setting of previous methods [5, 44] and is only trained on the RES datasets. As shown in Tab. 7, the proposed approach ReLA outperforms other methods on classic RES. Our performance is consistently higher than the state-of-the-art LAVT [44] with a margin of 1%~4% on three datasets. Although the performance gain of our proposed method over other methods on classic RES is lower than that on GRES, the results show that the explicit relationship modeling is also beneficial to classic RES. More results are reported in Supplementary Materials.

## 6. Conclusion

We analyze and address the limitations of the classic RES task, *i.e.*, it cannot handle multi-target and no-target expressions. Based on that, a new benchmark, called Generalized Referring Expression Segmentation (GRES), is defined to allow an arbitrary number of targets in the expressions. To support the research on GRES, we construct a large-scale dataset gRefCOCO. We propose a baseline method ReLA for GRES to explicitly model the relationship between different image regions and words, which consistently achieves new state-of-the-art results on the both classic RES and newly proposed GRES tasks. The proposed GRES greatly reduces the constraint to the natural language inputs, increases the application scope to the cases of multiple instances and no right objects in image, and opens possible new applications such as image retrieval.

# References

[1] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7454–7463, 2019. 3

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, volume 1, pages 4171–4186. Association for Computational Linguistics, 2019. 4

[4] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In *Proc. Eur. Conf. Comput. Vis.*, pages 417–435. Springer, 2020. 3

[5] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 16321–16330, 2021. 3, 7, 8

[6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 3, 8

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent.*, 2021. 4, 5, 6

[8] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 2006. 2

[9] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023. 3

[10] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023. 3

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[12] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1115–1124, 2017. 3

[13] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Proc. Eur. Conf. Comput. Vis.*, pages 108–124. Springer, 2016. 2

[14] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object

[15] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4424–4433, 2020. 3

[16] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10488–10497, 2020. 3

[17] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 59–75. Springer, 2020. 3

[18] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9858–9867, 2021. 7

[19] Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 2

[20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proc. of the Conf. on Empirical Methods in Natural Language Process.*, pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 1, 2, 3

[21] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 18145–18154, 2022. 5, 6, 8

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[23] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5745–5753, 2018. 3

[24] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10880–10889, 2020. 2

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[26] Chang Liu, Xudong Jiang, and Henghui Ding. Instance-specific feature propagation for referring segmentation. *IEEE Trans. Multimedia*, 2022. 3

[27] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4673–4682, 2019. 2

[28] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4856–4864, 2017. 3

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 4

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3431–3440, 2015. 3

[31] 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 Int. Conf. Multimedia*, pages 1274–1282, 2020. 3

[32] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10034–10043, 2020. 7, 8

[33] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7102–7111, 2017. 3

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 11–20, 2016. 1, 2

[35] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proc. Eur. Conf. Comput. Vis.*, pages 630–645, 2018. 3

[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[37] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7262–7272, 2021. 5

[38] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1960–1968, 2019. 2

[39] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 11686–11695, 2022. 3, 7, 8

[40] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10216–10225, 2020. 2, 6

[41] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 12077–12090, 2021. 5

[42] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Proc. Eur. Conf. Comput. Vis.*, volume 12359, pages 387–404. Springer, 2020. 2

[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 *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4683–4693, 2019. 2

[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 18155–18165, 2022. 3, 6, 7, 8

[45] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10502–10511, 2019. 3

[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1307–1315, 2018. 2, 3, 7

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proc. Eur. Conf. Comput. Vis.*, pages 69–85. Springer, 2016. 1, 2, 3

[48] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7282–7290, 2017. 3

[49] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6974–6983, 2021. 3

[50] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 557–566, 2017. 3

[51] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *arXiv preprint arXiv:2212.01769*, 2022. 3

[52] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4252–4261, 2018. 2