Improving Target Presence and Plurality Recognition for Generalized Referring Image Segmentation

Namyup Kim*, Jinsung Lee*, Suha Kwak
Pohang University of Science and Technology (POSTECH), Korea
{namyup, jinsunglee, suha.kwak}@postech.ac.kr

Abstract

Generalized referring image segmentation (RIS) aims to segment regions in an image described by a natural language expression, handling not only single-target but also no- and multi-target scenarios. Previous approaches have proposed new components that enable a conventional RIS model to handle these additional scenarios, such as a target presence prediction head for no-target scenarios and multiple mask candidates for multi-target cases. However, we observe that these methods predominantly rely on the conventional RIS backbone without fully integrating the additional components and thus still struggle in such general scenarios. To address this, we propose an effective framework specifically tailored to handle no-target and multi-target scenarios, incorporating both architectural and data-driven approaches. Our architecture employs a learnable query designed to understand both target presence and plurality. While this approach alone outperforms previous state-of-the-art methods with similar computational requirements, we further introduce a novel data augmentation strategy that enables our framework to surpass computationally intensive LMM-based approaches.

1 Introduction

Referring Image Segmentation (RIS) is the task of locating the region corresponding to a natural language description in an image (Liu et al. 2017; Li et al. 2018; Chen et al. 2019; Ye et al. 2019; Hu et al. 2020; Feng et al. 2021; Kim et al. 2022; Wang et al. 2022). Conventional RIS models and benchmarks are exclusively designed for segmenting only a single target. Due to this limitation, they fail to address diverse demands in real-world applications where a description may refer to multiple targets or those not present in the image. To resolve this issue, *generalized RIS* (Liu, Ding, and Jiang 2023) was introduced, which goes beyond the single-target scenario to handle additional multi-target and no-target situations as illustrated in Fig. 1. This new and more challenging RIS task requires a model to understand both whether the referenced object exists in the image and whether multiple objects are being referenced in the description.

ReLA (Liu, Ding, and Jiang 2023), the pioneering work on generalized RIS, incorporates an additional binary classification head to determine whether the target is present in the image and achieved superior to conventional RIS models. The target presence head, however, directly receives the same features used for mask prediction, forcing these features to serve dual purposes: segmentation and presence classification. This design can lead to conflicts, as the features may concentrate more on segmentation information, detracting from effective target presence classification (Wu et al. 2020; Song, Liu, and Wang 2020). Our experiments revealed that the performance of the target presence head of ReLA is far from optimal as the input features concentrate more on segmentation rather than the target presence (Sec. 4.5).

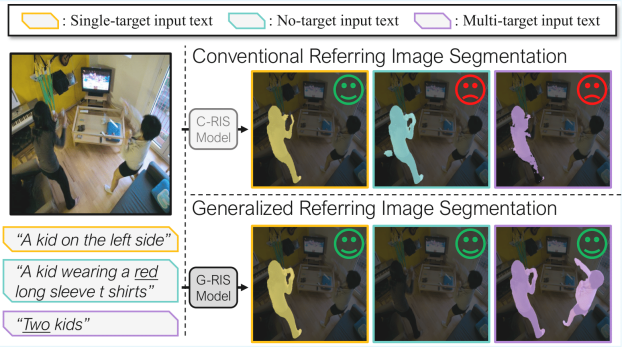

Figure 1: Comparison between conventional and generalized RIS models. While conventional RIS models are designed to identify only the most relevant region for a given text in single-target scenarios, generalized RIS models must handle more complex cases: determining whether the referenced object exists in the image (*target presence*) , and recognizing whether the text refers to single or multiple objects (*target plurality*). Our model explicitly addresses these challenges by improving both target presence and plurality recognition, leading to improved performance across all scenarios.

This work also introduces a new dataset for generalized RIS, which includes no- and multi-target samples for training and evaluation. Although valuable, adding such samples to a conventional RIS dataset increases annotation complexity and costs. Specifically, it takes greater annotation effort to generate non-trivial no-target expressions (*e.g.*, referring to "zebra" in an image containing only a giraffe) and diverse multi-instance clusters with corresponding textual descriptions (*e.g.*, everyone except the kid in white). Due to this issue, no-target and multi-target samples comprise a significantly smaller portion of the training set compared to single-target samples, potentially limiting the model's ability to handle these scenarios effectively.

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

To address the limitation of the target classification in the previous method, we propose a new architecture to improve target presence and plurality recognition among no-, single-, and multi-target cases. First, we introduce *target query* specifically designed for target presence classification. While the previous method (Liu, Ding, and Jiang 2023) extracts target presence information from the features used for mask prediction, our target query is explicitly designed to aggregate this information from the intermediate features of the pixel decoder (Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022; Liu, Ding, and Jiang 2023), which generates text-aligned masks by processing textual and visual features through multiple refinement layers. These intermediate features offer richer semantic context of the input scene than those used solely for mask prediction, enabling the target query to more effectively determine the target presence via cross-attention blocks (Dosovitskiy et al. 2021).

Moreover, to distinguish between single-target and multi-target scenarios, we attach a plurality classification head to the target query. Unlike previous methods (Liu, Ding, and Jiang 2023; Lai et al. 2024; Xia et al. 2024), our design ensures that the model captures plurality cues directly from text features. This distinct training signal not only enriches the target query with explicit plurality information but also offers a clear guidance for generating accurate masks under complex referring scenarios.

While these architectural changes alone bring clear improvements, we observe that the aforementioned severe data imbalance in the existing dataset (Liu, Ding, and Jiang 2023) limits the model to fully utilize the proposed design. Since the target query learns to identify target presence and plurality from no- and multi-target samples, it is crucial to provide sufficient instances of these scenarios during training. To address this, we introduce a new data augmentation strategy designed to amplify these underrepresented scenarios during training, allowing the model to better exploit its capacity for generalized target reasoning. In detail, we generate synthetic no-target samples by pairing input images with text descriptions that are originally coupled with other images within the same batch. For multi-target samples, we stitch multiple single-target images into one image and concatenate their captions to build a single long sentence describing multiple targets as synthetic multi-target labels. Despite the simplicity of this augmentation strategy, we empirically find it surprisingly effective in generating masks corresponding to complex referring scenarios.

In summary, our contribution is three-fold:

- We propose a new architecture that introduces *target query*, a learnable embedding specifically designed to improve generalized RIS through target presence classification and plurality classification.
- We develop a simple yet effective data augmentation strategy to efficiently generate synthetic no- and multi-target samples during training, thereby mitigating the data imbalance of the existing dataset for generalized RIS.
- Our model outperforms existing methods in terms of both efficacy and efficiency on all evaluation splits of the generalized RIS dataset.

2 Related Work

Referring Image Segmentation. Referring Image Segmentation (RIS) aims to segment image regions corresponding to natural language expressions. Early works used CNNs and RNNs for feature extraction (Hu, Rohrbach, and Darrell 2016; Liu et al. 2017), followed by the introduction of attention mechanisms and multi-level feature aggregation (Margffoy-Tuay et al. 2018; Chen et al. 2019; Ye et al. 2019; Hu et al. 2020). Recent transformer-based models (Kamath et al. 2021; Ding et al. 2021; Kim et al. 2022) and large-scale vision-language frameworks (Wang et al. 2022) have enabled more accurate segmentation. Generalized RIS (Liu, Ding, and Jiang 2023) expanded the task to include no-target and multi-target scenarios. Recent approaches leveraging large multimodal models (Lai et al. 2024; Xia et al. 2024) have shown notable improvements. Unlike these approaches that rely on complex architectures or large-scale pretraining, our method achieves state-of-the-art performance through a novel architecture for target presence and plurality recognition, combined with effective data augmentation.

Transformer-Based Semantic Segmentation. Following the success of CNNs that treat semantic segmentation as a per-pixel classification problem (Chen et al. 2015; Long, Shelhamer, and Darrell 2015; Noh, Hong, and Han 2015), the advent of vision transformers (Dosovitskiy et al. 2021) has sparked the development of transformer-based segmentation approaches (Li et al. 2023; Jain et al. 2023b,a; Cheng et al. 2022; Kirillov et al. 2023), which have demonstrated superior performance in this task. Among them, MaskFormer (Cheng, Schwing, and Kirillov 2021) stands out for reframing segmentation as a mask classification task using a DETR-style encoder-decoder architecture (Carion et al. 2020), eliminating the need for computationally heavy per-pixel predictions. Our method similarly employs the encoder-decoder architecture with multi-scale deformable attention (Zhu et al. 2020) in its decoder, thereby directly producing masks without requiring multiple mask proposals.

Data Augmentation for Generalization. Data augmentation is a technique to generate new data samples using existing datasets, which aims to increase model's generalization capability. Although a key strategy for successful data augmentation has been a modification of existing samples in a way that does not affect its label prediction (Wang, Perez et al. 2017; Simonyan and Zisserman 2014; Yun et al. 2019; Zhong et al. 2020), the studies that require negative samples such as generative adversarial network (Goodfellow et al. 2020), metric learning (Hadsell, Chopra, and LeCun 2006), or contrastive learning (Melekhov, Kannala, and Rahtu 2016) have developed the augmentation technique to artificially generate false samples (Sinha et al. 2021; Wang and Qi 2022; Duan et al. 2018). These methods introduce mismatched samples to help models learn negative relationships, promoting a more balanced understanding of real-world scenarios. Inspired by this, we augment no-target cases using image-text pairs without correspondence, encouraging the model to recognize misalignment. We further construct multi-target examples by combining images and captions from different samples, exposing the model to diverse referential scenarios.

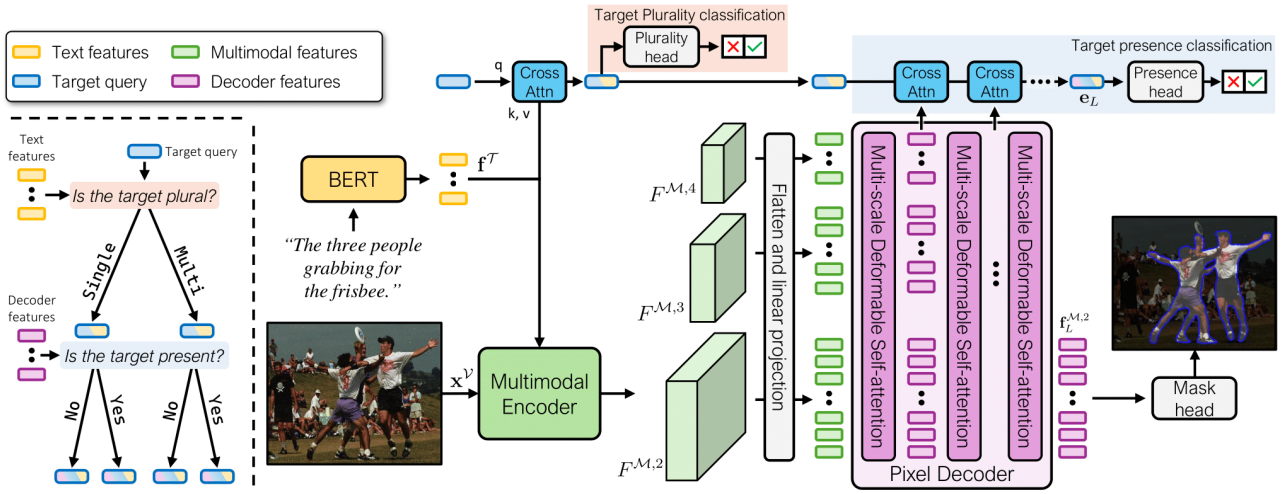

Figure 2: Overview of the proposed architecture. Given an image and text pair as input, our framework consists of: a multimodal encoder that computes hierarchical multimodal feature maps $F^{\mathcal{M},i}$, a pixel decoder that processes these features to generate mask predictions $\mathbf{f}_L^{\mathcal{M},2}$, a plurality head that provides a learning signal to classify whether the input text refers to single or multiple targets, and a presence head that performs a prediction on whether the referenced objects are present in the image. The left diagram illustrates how features are progressively merged into the target query.

3 Proposed Method

This section elaborates on our framework for generalized Referring Image Segmentation (RIS). We first describe the overall model architecture (Sec. 3.1) and then present how the proposed target query performs target presence classification and target plurality classification (Sec. 3.2). Moreover, we explain our data augmentation strategy for no- and multi-target samples to ensure a sufficient and diverse set of both target-type samples (Sec. 3.3). Finally, we provide details on the training and inference processes (Sec. 3.4).

3.1 Overall Architecture

The overall architecture of our framework is illustrated in Fig. 2. It consists of feature extractors, a pixel decoder, a presence head, and a plurality head. The feature extractors, consisting of a text encoder and a multimodal encoder, compute multimodal feature maps, which are then fed into the pixel decoder to produce a segmentation mask prediction. The plurality head utilizes text features to perform plurality classification, whereas the presence head conducts target presence classification using the decoder features as input.

Feature Extractors Given an image and text pair (I, T) as inputs, the image is split into non-overlapping patches and linearly projected to obtain patch embeddings $\mathbf{x}^{\mathcal{V}} \in \mathbb{R}^{N_p \times d_v}$, where N_p denotes the number of patches. Concurrently, the text is encoded into text features $\mathbf{f}^{\mathcal{T}} \in \mathbb{R}^{N_t \times d_t}$ using a transformer-based text encoder (Devlin et al. 2019), with N_t representing the number of words. We then adopt Swin Transformer (Liu et al. 2021) as the multimodal encoder on top of the patch embeddings and text features, as in (Yang et al. 2022; Liu, Ding, and Jiang 2023). The multimodal encoder $\Phi(\cdot,\cdot)$ takes the patch embeddings $\mathbf{x}^{\mathcal{V}}$ and the text features $\mathbf{f}^{\mathcal{T}}$ to produce a set of multi-scale feature maps:

$$\Phi(\mathbf{x}^{\mathcal{V}}, \mathbf{f}^{\mathcal{T}}) = \left\{ F^{\mathcal{M},i} \mid F^{\mathcal{M},i} \in \mathbb{R}^{H_i \times W_i \times C},\ i \in [1,4] \right\} \tag{1}$$

where $F^{\mathcal{M},i}$ is the feature extracted from the i-th layer of the backbone, and (H_i, W_i) is its spatial dimension.

Pixel Decoder Following Mask2Former (Cheng et al. 2022), the pixel decoder $\Psi(\cdot)$ consists of an L series of Multi-Scale Deformable Self-attention (MSDS) blocks where the l-th MSDS block is denoted as $\texttt{MSDS}_l(\cdot)$. Each MSDS block operates similarly to a Transformer layer (Vaswani et al. 2017), processing a sequence of inputs and producing a refined output of the same shape via an attention mechanism. However, instead of computing pairwise interactions between all input elements, it attends to a small set of dynamically sampled key positions at multiple feature scales. Thus, the output of the MSDS block captures the semantic relationship among input elements, while avoiding the quadratic computational cost of standard self-attention. Due to the space constraint, we omit the detailed mechanism of the MSDS block and refer readers to the original implementation (Cheng et al. 2022; Zhu et al. 2020) for further details.

We flatten multi-scale feature maps $\{F^{\mathcal{M},j}\}_{j=2}^{4}$ from the multimodal encoder and apply linear projection to obtain $\{\mathbf{f}_0^{\mathcal{M},j} \mid \mathbf{f}_0^{\mathcal{M},j} \in \mathbb{R}^{H_j W_j \times C},\ j \in [2,4]\}$. Then, these flattened feature maps are concatenated along the flattened dimension to form a unified sequence of shape $\left(\sum_{j=2}^{4} H_j W_j, C\right)$, and processed with L MSDS blocks:

$$\begin{aligned} \Psi &= \texttt{MSDS}_L \circ \texttt{MSDS}_{L-1} \circ \cdots \circ \texttt{MSDS}_1, \\ \mathbf{f}_l^{\mathcal{M}} &= \texttt{MSDS}_l(\mathbf{f}_{l-1}^{\mathcal{M}}), \end{aligned} \tag{2}$$

where $\mathbf{f}_l^{\mathcal{M}} = [\mathbf{f}_l^{\mathcal{M},2}, \mathbf{f}_l^{\mathcal{M},3}, \mathbf{f}_l^{\mathcal{M},4}]$ and $[\cdot,\cdot]$ denotes concatenation. The output $\mathbf{f}_L^{\mathcal{M}}$ is then passed to the mask head to generate the mask prediction. Specifically, to obtain the fine-grained mask prediction, the highest-resolution feature $\mathbf{f}_L^{\mathcal{M},2}$ from $\mathbf{f}_L^{\mathcal{M}}$ is reshaped to its original spatial size (H_2, W_2) and upsampled via interpolation to match the spatial size of $F^{\mathcal{M},1}$. The two feature maps are then summed, upscaled

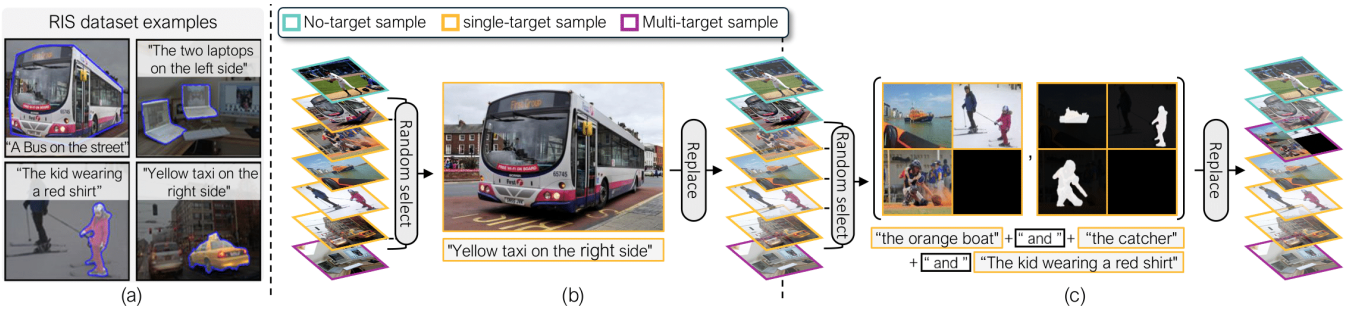

Figure 3: Illustration of our data augmentation for generating no-target and multi-target samples. (a) RIS datasets examples: Since each sample has a distinctive and detailed text caption, its description hardly refers to any valid region in other data samples. (b) No-target sample generation: We randomly select a single-target sample and replace its text with another text from within the batch while removing the ground-truth mask. (c) Multi-target sample generation. Among the remaining single-target samples, we randomly select up to 4 samples and arrange them in a 2×2 grid. Their corresponding texts are concatenated with " and " to create a multi-target sample.

with interpolation again to match the input spatial dimension (H, W), and passed through a Multi-Layer Perception (MLP) to generate the final mask prediction $\hat{Y} \in \mathbb{R}^{H \times W \times 2}$.

Presence and Plurality Heads Both the presence and plurality heads play a crucial role in our architecture by directly guiding the model to identify the scenario it encounters–whether the target is present in the image and whether it refers to a single instance or multiple instances (left diagram of Fig. 2). Each head consists of a single MLP that takes as input the target query, whose design and supervision will be detailed in the next section, and produces a binary prediction: the presence head outputs $\hat{y}_{\text{pr}} \in \mathbb{R}^2$ for target presence, while the plurality head outputs $\hat{y}_{\text{pl}} \in \mathbb{R}^2$ for target plurality.

3.2 Plurality and Presence Classification

We deliver the information about the underlying referring scenario of the input to the model using a learnable query, referred to as the *target query*, which serves as an essential component for guiding the encoder and decoder toward a generalized understanding of the visual-textual context. We begin by describing how it encodes information to determine plurality of the referred instance, and then explain its strategy for identifying the target presence.

Plurality Classification In referring image segmentation, user intent is conveyed via text prompts, which inherently carry cues about the plurality of the referred instances. For example, the text prompt "*The three people grabbing for the frisbee.*" from Fig. 2 alone reveals that the model needs to capture multiple instances in the input image. In fact, our empirical findings show that textual cues are more than enough to accurately determine the plurality of the referred instance (Sec. 4.5, Fig. 5(b)). Accordingly, we let the target query attend to the text features $\mathbf{f}^{\mathcal{T}}$ and classify whether there are multiple instances to refer to. Specifically, we initialize the target query $\mathbf{e}'_0 \in \mathbb{R}^C$ and use it as the query in a cross-attention block, with the text features $\mathbf{f}^{\mathcal{T}}$ serving as both the key and value:

$$\mathbf{e}_0 = \texttt{CrossAttn}(\mathbf{f}^{\mathcal{T}}; \mathbf{e}'_0) \in \mathbb{R}^C. \tag{3}$$

The resulting target query $\mathbf{e}_0$ is then fed into the plurality head to generate the target plurality prediction $\hat{y}_{\text{pl}}$.

Presence Classification In contrast to the plurality classification, which relies primarily on textual cues, presence classification requires the model to jointly understand both the textual and visual context. Since the target query $\mathbf{e}_0$ already encodes the textual information, we reuse it to integrate visual cues for determining the target presence. To this end, we extract visual features from intermediate features of the pixel decoder as they provide rich, multi-level semantic information, in contrast to prior work (Liu, Ding, and Jiang 2023) that relies solely on the final pixel decoder output for the target presence prediction. Specifically, the target query $\mathbf{e}_0$ aggregates information required for target presence classification from $\{\mathbf{f}_l^{\mathcal{M}}\}_{l=1}^{L}$, using cross-attention blocks (Vaswani et al. 2017) as follows:

$$\mathbf{e}_l = \texttt{CrossAttn}_l(\mathbf{f}_{l-1}^{\mathcal{M},4}; \mathbf{e}_{l-1}) \in \mathbb{R}^C. \tag{4}$$

Note that we utilize only the subsequence $\mathbf{f}_l^{\mathcal{M},4}$ from the flattened intermediate feature $\mathbf{f}_l^{\mathcal{M}} = [\mathbf{f}_l^{\mathcal{M},2}, \mathbf{f}_l^{\mathcal{M},3}, \mathbf{f}_l^{\mathcal{M},4}]$ to avoid excessive computational and memory overhead. Since the subsequence $\mathbf{f}_l^{\mathcal{M},4}$ provides higher-level semantic context while having the shortest length among others, it enables both computationally efficient and semantically rich target presence classification. Supplementary results in Sec. A show that this choice leads to the best performance, while using the whole intermediate feature results in memory outage. Finally, we feed the output target query $\mathbf{e}_L$ to the presence head to produce the target presence prediction $\hat{y}_{\text{pr}}$.

3.3 Data Augmentation for Generalized RIS

To fully leverage our architecture's ability to handle diverse referring scenarios, including cases where multiple or no target instances are referred, it is crucial to expose the model to a broader spectrum of such examples during training. However, we figure that the data imbalance within existing benchmarks poses a challenge: no-target samples account for only 9.14% of the dataset, while multi-target samples make up 25.53%, leaving the majority of the data dominated by traditional single-target cases. This imbalanced training sample distribution limits the effectiveness of training the target query to accurately recognize diverse referring scenarios, thereby

constraining the full potential of our model architecture. To address this, we introduce a data augmentation strategy that generates additional no-target and multi-target samples.

Our data augmentation strategy transforms a specified proportion of single-target samples within each batch into no- and multi-target samples. Given a training batch $B = \{(I_i, T_i, Y_i)\}_{i=1}^{N_b}$, where N_b is the batch size and each triplet (I_i, T_i, Y_i) represents the i-th training sample consisting of an image, a text query, and the corresponding ground-truth mask, respectively, we collect all N_s single-target samples in the batch to construct $\tilde{B} = \{(I_j, T_j, Y_j)\}_{j=1}^{N_s}$, which serves as the source samples for constructing synthetic no- and multi-target samples.

Generating Synthetic No-Target Samples Our strategy for no-target sample generation leverages a key property of RIS datasets: the text descriptions are often highly detailed and tailored to unambiguously identify specific regions within their paired images. Due to its specificity, a caption rarely applies meaningfully to other images; for example, the caption "*Yellow taxi on the right side*" shown in Fig. 3(a) is highly unlikely to correspond to any valid region in different images, as its referential scope is too narrow and context-dependent. Hence, we generate no-target samples by replacing a description of a sample in $\tilde{B}$ with a description from a different sample in $\tilde{B}$, with a probability of τ_{no}. While this strategy may occasionally produce trivial no-target samples (*e.g.*, referring to a "cake" in an image full of elephants), it remains a simple and efficient method for generating diverse no-target scenarios, providing useful learning signals for enhancing target presence classification.

Generating Synthetic Multi-Target Samples To generate multi-target samples, we randomly select up to 4 single-target samples from $\tilde{B}$, with each sample being selected independently with a probability τ_{mul}. As illustrated in Fig. 3(c), we construct an augmented image by placing images from selected samples into random cells of a 2×2 grid and generate the corresponding mask by positioning each sample's mask in its respective cell. The accompanying text description is created by concatenating the individual descriptions in grid order using the conjunction "and". The resulting synthetic triplet then replaces one of the 4 chosen single-target samples used for augmentation, preserving the batch size while balancing the proportion of different referring scenarios within the batch. Although this strategy generates multi-target samples that may overlook directional clues in the referring descriptions, it effectively help the model learn to segment multiple instances with diverse attributes, thereby improving the mask prediction quality.

3.4 Training and Inference

Our model is trained with three different losses: pixel-wise cross-entropy loss $\mathcal{L}_{\text{pixel}}(\hat{Y}, Y)$, presence loss $\mathcal{L}_{\text{ce}}^{\text{pr}}(\hat{y}_{\text{pr}}, y_{\text{pr}})$, and plurality loss $\mathcal{L}_{\text{ce}}^{\text{pl}}(\hat{y}_{\text{pl}}, y_{\text{pl}})$. Here, the ground-truth presence label y_{pr} can be directly inferred from standard generalized RIS annotations; it is set to 1 when $Y \neq \mathbf{0}$, and 0 otherwise. In contrast, the ground-truth plurality label may appear unavailable at first glance, but it is actually embedded in the original annotations: since the datasets are constructed by combining multiple masks, this information is freely available, although often discarded by prior methods. Combining all together, the total loss is given by

$$\mathcal{L}_{total} = \mathcal{L}_{\text{pixel}} + \lambda_{\text{pr}}\mathcal{L}_{\text{ce}}^{\text{pr}}(\hat{y}_{\text{pr}}, y_{\text{pr}}) + \lambda_{\text{pl}}\mathcal{L}_{\text{ce}}^{\text{pl}}(\hat{y}_{\text{pl}}, y_{\text{pl}}), \tag{5}$$

where λ_{pr} and λ_{pl} are balancing hyperparameters for their respective losses. During inference, we obtain the binary mask prediction by thresholding each region of the predicted mask probability $\hat{Y}$ by 0.5. The target presence prediction $\hat{y}_{\text{pl}}$ is also used to refine the final output: if the predicted presence probability is below 0.5, our model disregards the binary mask and outputs no segmentation mask at all.

4 Experiments

In this section, we present a detailed experimental setting, including the datasets and metrics (Sec. 4.1), as generalized RIS has recently been introduced. Then, we provide the implementation details of our framework (Sec. 4.2). To demonstrate its effectiveness, we compare our framework with the existing RIS methods on the generalized RIS dataset (Sec. 4.3). Following this, we conduct ablation study (Sec. 4.4) and various in-depth analyses (Sec. 4.5).

4.1 Datasets and Metrics

gRefCOCO. We primarily conduct experiments on the generalized referring image segmentation dataset gRefCOCO (Liu, Ding, and Jiang 2023). which consists of 19,994 images and 259,859 expressions for 83,513 masks. As shown in Tab. A4 of the supplement, the proportion of no-target samples in the *train* set is significantly lower than in the evaluation splits, emphasizing the chllenge of training the model under limited data conditions. As mentioned earlier in Sec. 3.4, the multi-target samples in this dataset are labeled based on COCO dataset's instance segmentation indices, which automatically provide plurality labels for the corresponding expressions.

gIoU, cIoU, and N-acc. Following the recent approach in the literature (Liu, Ding, and Jiang 2023), we adopt the generalized Intersection-over-Union (gIoU) metric, which extends the standard mean IoU by incorporating no-target cases. While the standard mean IoU is calculated by first computing IoU for each image and then averaging across all evaluation samples, gIoU further accounts for scenarios where the target is absent: it assigns an IoU of 1 for a true positive no-target sample and 0 otherwise, making it a more comprehensive metric under generalized referring segmentation settings. We also compute the cumulative IoU (cIoU), where total intersections are divided by the total unions over all evaluation samples. Note that cIoU is less sensitive to the true positive number of no-target samples compared to gIoU, and weighs more on accurate prediction of single-target and multi-target samples. We also utilize the metric that computes the recall of no-target samples to measure the presence classification performance: N-acc.$= \frac{\text{TP}_{\text{NO}}}{\text{TP}_{\text{NO}}+\text{FN}_{\text{NO}}}$, where TP_{NO} and FN_{NO} denote the number of true positive and false negative of no-target samples, respectively. Since the evaluation

Methods	Vision Backbone (#param.)	val			testA			testB		
		gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.
MattNet (Yu et al. 2018)	ResNet-101 (44.2M)	48.24	47.51	41.15	59.30	58.66	44.04	46.14	45.33	41.32
LTS (Jing et al. 2021)	DarkNet-53 (41M)	52.70	52.30	-	62.64	61.87	-	50.42	49.96	-
VLT (Ding et al. 2021)	DarkNet-53 (41M)	52.00	52.51	47.17	63.20	62.19	48.74	50.88	50.52	47.82
CRIS (Wang et al. 2022)	CLIP-R101 (44.2M)	56.27	55.34	-	63.42	63.82	-	51.79	51.04	-
LAVT (Yang et al. 2022)	Swin-B (87.7M)	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA (Liu, Ding, and Jiang 2023)	Swin-B (87.7M)	63.60	62.42	56.37	70.03	69.26	59.02	61.02	59.88	58.40
LISA (Lai et al. 2024)†	SAM-H (240M)	61.63	61.76	54.67	66.27	68.50	50.01	58.84	**60.63**	51.91
GSVA (Xia et al. 2024)†	SAM-H (240M)	66.47	63.29	62.43	71.08	69.93	65.31	62.23	60.47	60.56
Ours	Swin-B (87.7M)	**71.32**	**66.61**	**72.27**	**72.54**	**70.55**	**70.61**	**62.24**	60.01	**61.74**

Table 1: Comparison with existing methods on the gRefCOCO dataset, where † indicates methods that utilize the LMM approach with LLaVa-Vicuna-7B (Liu et al. 2024).

on models without a target presence prediction head is not straightforward, we follow prior work (Liu, Ding, and Jiang 2023) and adopt a threshold-based heuristic: if the predicted binary mask contains fewer than 50 pixels (about 0.02% of the image size), the target is considered absent; otherwise, it is considered present.

4.2 Implementation Details

Following previous works (Yang et al. 2022; Liu, Ding, and Jiang 2023), we use Swin-Base (Liu et al. 2021) pre-trained on ImageNet-22K as our visual backbone and BERT-base (Devlin et al. 2019) as our text encoder. The pixel decoder consists of 6 MSDS blocks with 256 channels. During training, we resize input images to 480×480 pixels while maintaining their aspect ratio through padding. We use AdamW optimizer (Loshchilov and Hutter 2019) with a weight decay of 0.01 and an initial learning rate of 2.5×10^{-6} for the backbone networks and 2.5×10^{-5} for other components. The learning rate follows a polynomial decay schedule with a power of 0.9. We train our model for 150,000 iterations with a batch size of 48 using PyTorch on 2 NVIDIA A6000ada GPUs. For data augmentation, we set both probabilities τ_{no} and τ_{mul} to 0.05. The loss weights λ_{pr} and λ_{pl} are set to 0.1 and 0.01, respectively. Our implementation will be made publicly available.

4.3 Comparison with Previous Arts

We evaluate our framework on the gRefCOCO dataset and compare it with state-of-the-art methods in RIS, as summarized in Tab. 1. Our model outperforms existing approaches, including recent Large Multimodal Models (LMMs) such as LISA (Lai et al. 2024) and GSVA (Xia et al. 2024). Despite their use of advanced backbone networks (Zou et al. 2023) with significantly more parameters, our model achieves superior performance in a more parameter-efficient manner.

Notably, the improvement from our data augmentation strategy is orthogonal to the model architecture and can be applied to other methods as well. To demonstrate this, we quantify its impact on the previous state of the art, ReLA (Liu, Ding, and Jiang 2023), and compare its performance with ours under the same conditions in Fig. 4. Note that gIoU measures overall performance, cIoU mainly assesses the single-target and multi-target mask prediction quality, and N-acc evaluates target presence classification. Despite the concerns

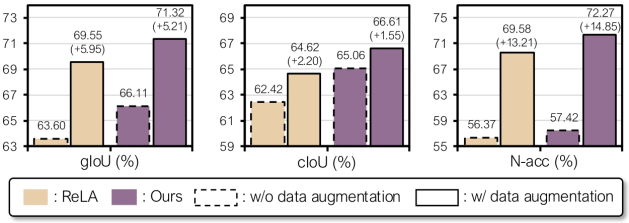

Figure 4: Performance comparison between ReLA and our model on the gRefCOCO *val* set with and without data augmentation.

Target Query	Plurality Cls.	DA for No	DA for Multi	gIoU	cIoU	N-acc.
				62.39	59.62	54.97
✓				64.93	63.86	55.60
✓	✓			66.11	65.05	57.42
✓	✓	✓		69.24	63.61	71.28
✓	✓	✓	✓	**71.32**	**66.61**	**72.27**

Table 2: Ablation study on the *val* split of gRefCOCO dataset. The results demonstrate that each component—target query for target classification, plurality classification, and data augmentation (DA) for no- and multi-target samples—contributes significantly to overall performance improvements across all metrics. The best results are achieved when all components are combined.

raised in Sec. 3.3, our data augmentation consistently improves all metrics regardless of the model architecture, indicating its effectiveness in enhancing both segmentation mask quality and target presence classification. Moreover, the performance of our model outperforms ReLA even without the data augmentation strategy, implying that the proposed architecture alone can offer a performance advantage over existing approaches. In summary, these results highlight two key conclusions: (1) the proposed augmentation strategy is compatible with other generalized RIS models, making it broadly applicable; and (2) our model achieves strong performance on its own, validating the effectiveness of its architecture.

4.4 Ablation Study

As summarized in Tab. 2, we ablate all of our proposed methods, including the target query guided by the target presence classification, the target plurality classification, and the application of the data augmentation strategies.

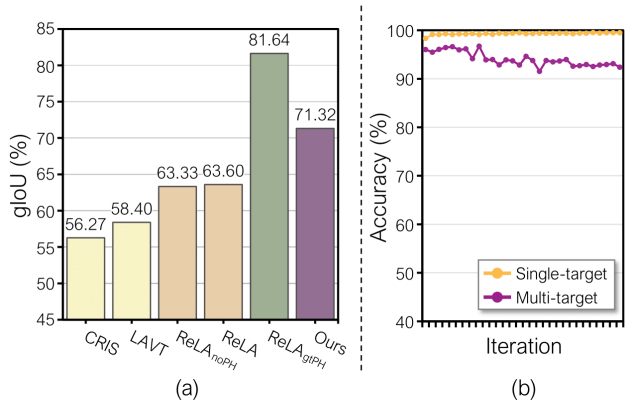

Figure 5: (a) Performance on the gRefCOCO *val* set in gIoU (%). ReLA$_{\text{noPH}}$ denotes ReLA without the Presence Head (PH), which only uses the mask prediction head to output final prediction. ReLA$_{\text{gtPH}}$ represents ReLA refined with ground-truth presence label, indicating the optimal performance when its presence head works perfectly. (b) Plurality classification accuracy during training on the gRefCOCO *val* set, shown separately for single- and multi-target samples.

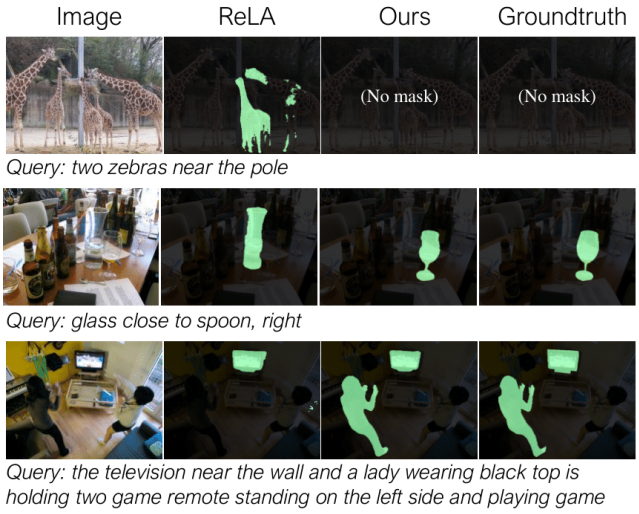

Figure 6: Qualitative results of ReLA (Liu, Ding, and Jiang 2023) and Ours on gRefCOCO *val* set. Each row shows no-, single-, and multi-target text query results.

Target Query The target query alone improves the model's performance significantly across all metrics. Specifically, it increases gIoU and cIoU by 2.54%p and 4.24%p, respectively. This demonstrates that our dedicated target query effectively handles target presence classification while maintaining strong mask prediction performance.

Target Plurality Classification Adding plurality classification further enhances the model's capabilities, with improvements of 1.18%p in gIoU, 1.19%p in cIoU, and 1.82%p in N-acc. These results suggest that incorporating plurality information delivers the model helpful information generating the accurate mask prediction.

Data Augmentation for Generalized RIS Our data augmentation strategies contribute significantly to model performance improvements. Applying only the no-target sample augmentation already results in substantial gains, with improvements of 3.13%p in gIoU and 13.86%p in N-acc. Adding multi-target sample augmentation further boosts the performance on every metric, compensating for the slight drop in cIoU introduced by no-target augmentation alone. This suggests that generating only no-target samples is insufficient to cover the diverse scenarios of generalized RIS, while adding multi-target cases can lead to balanced performance improvements.

4.5 In-depth Analysis

Impact of Presence Head Quality Our analysis reveals that ReLA's presence head exhibits insufficient performance, suggesting that it fails to fulfill its intended role effectively, as demonstrated in Fig. 5(a). While the attachment of the presence head in ReLA brings a marginal performance of only 0.27%p (ReLA vs. ReLA$_{\text{noPH}}$), replacing its target presence prediction with ground-truth labels (ReLA$_{\text{gtPH}}$) leads to a dramatically gain of 18.04%p, indicating that the performance bottleneck lies in the accuracy of the target presence classification. Our model, on the other hand, demonstrates a significant performance gap between ReLA (71.32 vs. 63.60), primarily driven by a large discrepancy in the presence classification metric N-acc. (72.27 vs 56.37) as shown in Tab. 1.

Feasibility of Target Plurality Classification As shown in Fig. 5(b), our model achieves high accuracy in plurality classification from the early stages of training for both single- and multi-target samples. The plurality classification accuracy quickly reaches over 90% for both single- and multi-target samples, maintaining stable performance throughout training. This shows that plurality information can be effectively extracted from text features alone.

Qualitative Analysis As illustrated in Fig. 6, our model demonstrates strong segmentation performance across diverse referring expressions, handling no-, single-, and multi-target scenarios, effectively.

5 Conclusion

In this work, we presented a new framework for generalized referring image segmentation that effectively handles no-, single-, and multi-target scenarios. Our approach introduces a target query specifically designed for target presence classification, along with a plurality classification that enhances the model's understanding of target multiplicity. To address the data scarcity issue in no- and multi-target scenarios, we also propose a simple yet effective data augmentation strategy that generates synthetic samples during training. While our work makes significant progress, there remain opportunities for future research. For instance, exploring more sophisticated data augmentation techniques or investigating ways to better handle complex spatial relationships in multi-target scenarios could further advance the field.

Acknowledgements

This work was supported by the NRF grant and the IITP grant funded by Ministry of Science and ICT, Korea (NRF-2021R1A2C3012728, RS-2024-00509258, RS-2024-00469482, RS-2021-II212068, RS-2019-II191906).

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision (ECCV)*.

Chen, D.-J.; Jia, S.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019. See-through-text grouping for referring image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2015. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *Proc. International Conference on Learning Representations (ICLR)*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference ofthe North American Chapter ofthe Association for Computational Linguistics*.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-language transformer and query generation for referring segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*.

Duan, Y.; Zheng, W.; Lin, X.; Lu, J.; and Zhou, J. 2018. Deep adversarial metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Feng, G.; Hu, Z.; Zhang, L.; and Lu, H. 2021. Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *Proc. European Conference on Computer Vision (ECCV)*.

Hu, Z.; Feng, G.; Sun, J.; Zhang, L.; and Lu, H. 2020. Bi-directional relationship inferring network for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023a. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2989–2998.

Jain, J.; Singh, A.; Orlov, N.; Huang, Z.; Li, J.; Walton, S.; and Shi, H. 2023b. Semask: Semantically masked transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 752–761.

Jing, Y.; Kong, T.; Wang, W.; Wang, L.; Li, L.; and Tan, T. 2021. Locate then segment: A strong pipeline for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multimodal understanding. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Kim, N.; Kim, D.; Lan, C.; Zeng, W.; and Kwak, S. 2022. Restr: Convolution-free referring image segmentation using transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9579–9589.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3041–3050.

Li, R.; Li, K.; Kuo, Y.-C.; Shu, M.; Qi, X.; Shen, X.; and Jia, J. 2018. Referring image segmentation via recurrent refinement networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, C.; Ding, H.; and Jiang, X. 2023. Gres: Generalized referring expression segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 23592–23601.

Liu, C.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; and Yuille, A. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Proc. Neural Information Processing Systems (NeurIPS)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. *Proc. International Conference on Learning Representations (ICLR)*.

Margffoy-Tuay, E.; Pérez, J. C.; Botero, E.; and Arbeláez, P. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *Proc. European Conference on Computer Vision (ECCV)*.

Melekhov, I.; Kannala, J.; and Rahtu, E. 2016. Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*.

Noh, H.; Hong, S.; and Han, B. 2015. Learning Deconvolution Network for Semantic Segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinha, A.; Ayush, K.; Song, J.; Uzkent, B.; Jin, H.; and Ermon, S. 2021. Negative data augmentation. *arXiv preprint arXiv:2102.05113*.

Song, G.; Liu, Y.; and Wang, X. 2020. Revisiting the sibling head in object detector. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. Neural Information Processing Systems (NeurIPS)*.

Wang, J.; Perez, L.; et al. 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*.

Wang, X.; and Qi, G.-J. 2022. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. Cris: Clip-driven referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; and Fu, Y. 2020. Rethinking classification and localization for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3858–3869.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-modal self-attention network for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proc. International Conference on Learning Representations (ICLR)*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36: 19769–19782.