MaskRIS: Semantic Distortion-aware Data Augmentation for Referring Image Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Referring Image Segmentation (RIS) is an advanced vision-language task that involves identifying and segmenting objects within an image as described by free-form text descriptions. While previous studies focused on aligning visual and language features, exploring training techniques, such as data augmentation, remains underexplored. In this work, we explore effective data augmentation for RIS and propose a novel training framework called Masked Referring Image Segmentation (MaskRIS). We observe that the conventional image augmentations fall short of RIS, leading to performance degradation, while simple random masking significantly enhances the performance of RIS. MaskRIS uses both image and text masking, followed by Distortion-aware Contextual Learning (DCL) to fully exploit the benefits of the masking strategy. This approach can improve the model's robustness to occlusions, incomplete information, and various linguistic complexities, resulting in a significant performance improvement. Experiments demonstrate that MaskRIS can easily be applied to various RIS models, outperforming existing methods in both fully supervised and weakly supervised settings. Finally, MaskRIS achieves new state-of-the-art performance on RefCOCO, RefCOCO+, and RefCOCOg datasets.

1 Introduction

Referring Image Segmentation (RIS) (Hu et al., 2016) involves precisely delineating objects within an image based on text descriptions. Unlike semantic and instance segmentation, which are constrained to pre-defined classes, RIS offers unique flexibility by segmenting objects specified by free-form expressions. This capability has led to extensive applications in diverse domains, such as language-driven human-robot interaction (Shah et al., 2023) and advanced image editing (Chen et al., 2018; Patashnik et al., 2021). One of the main challenges in RIS is effectively bridging the modality gap between visual and language features. Recent research (Liu et al., 2023c; Yang et al., 2022; Wang et al., 2022) has focused on developing sophisticated architectures for cross-modal alignment.

While significant progress has been made in aligning different modalities, effective training techniques for RIS remain underexplored. The trend toward developing larger and more complex models requires challenging training settings like data augmentation. Previous studies (Cubuk et al., 2018; 2020; Touvron et al., 2021) have demonstrated the crucial role of data augmentation in model training. However, designing data augmentation for RIS is challenging: standard augmentation used in semantic segmentation is often incompatible with RIS, as some textual descriptions directly conflict with augmentation.

As illustrated in Figure 1, spatial augmentation, such as random crop and horizontal flip, alters the meaning of "left", causing the target mask to become incorrect after augmentation. Similarly, color distortion also conflicts with color-specific descriptions like "red". Due to the incompatibilities, naïvely applying augmentations is not beneficial - our empirical analysis reveals that conventional data augmentation degrades RIS (Liu et al., 2023c) performance (Figure 3(a)). This finding explains why previous RISs (Kim et al., 2022; Wang et al., 2022; Yang et al., 2022; Liu et al., 2023c; Wu et al., 2023; Zhang et al., 2022) have taken a cautious approach to augmentation, predominantly relying on simple resizing while avoiding complex techniques that could severely distort the data.

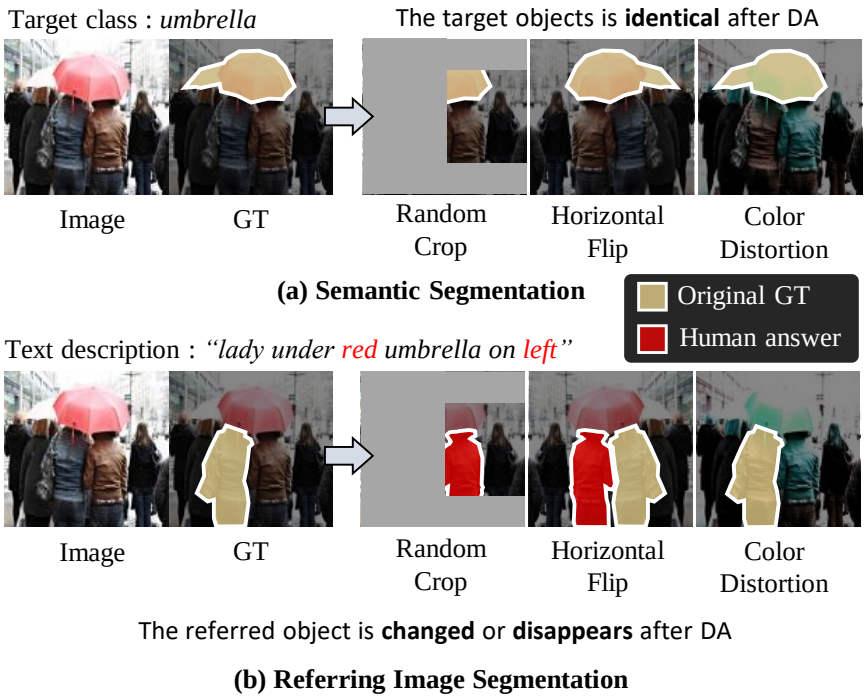


Figure 1: Conventional data augmentations (DA) in semantic segmentation are incompatible with referring image segmentation. Random crop and horizontal flip could change the referred object (*e.g.*, *"lady under the red umbrella on left"*) to another one, and color distortion could make the described object disappear.

In this paper, we explore an effective data augmentation framework to improve RIS model training. We propose a straightforward yet robust baseline, coined Masked Referring Image Segmentation (MaskRIS), which consists of two key components: (1) input masking as a data augmentation technique for RIS and (2) a Distortion-aware Contextual Learning (DCL) designed to maximize the benefits of input masking. In RIS tasks, it is crucial to preserve essential spatial information (*e.g.*, relative positions, ordering) and key attribute details (*e.g.*, color) for accurately understanding and segmenting objects based on referring expressions. Unlike conventional data augmentation, which often distorts these critical aspects, input masking minimizes such distortions by preserving spatial and attribute information while significantly expanding data diversity. Our approach extends input masking to both images and text to strengthen the model's capability to handle visual and linguistic complexity. Masking parts of the text encourages the model to infer missing or ambiguous details, improving its understanding of diverse referring expressions and reducing reliance on specific terms. This masking strategy directly addresses the need for a robust RIS training framework, helping to overcome the performance bottleneck caused by limited augmentation.

To fully exploit the advantages of input masking, we propose a Distortion-aware Contextual Learning (DCL) framework that processes inputs through two complementary paths: a primary path with original inputs and a secondary path with masked inputs. The primary path ensures baseline training stability and maintains model accuracy, while the secondary path enhances data diversity and robustness. The distillation loss aligns predictions from both paths, encouraging the model to make consistent predictions on both original and masked inputs. This DCL framework enhances feature robustness to masked inputs through the secondary path, while the primary path provides stability and mitigates potential harmful effects of severe distortion. Additionally, DCL acts as a regularizer, adding beneficial challenges to the RIS task. The regularization effect aligns with findings from Mobahi et al. (2020), where self-distillation is shown to constrain model representations, reducing overfitting and enhancing robustness - a process similarly achieved by aligning original and masked predictions in our framework.

Finally, we note that although MaskRIS shares certain surface-level similarities with the concurrent work CM-MaskSD (Wang et al., 2024), it is conceptually and technically distinct. As illustrated in Figure 2, CM-

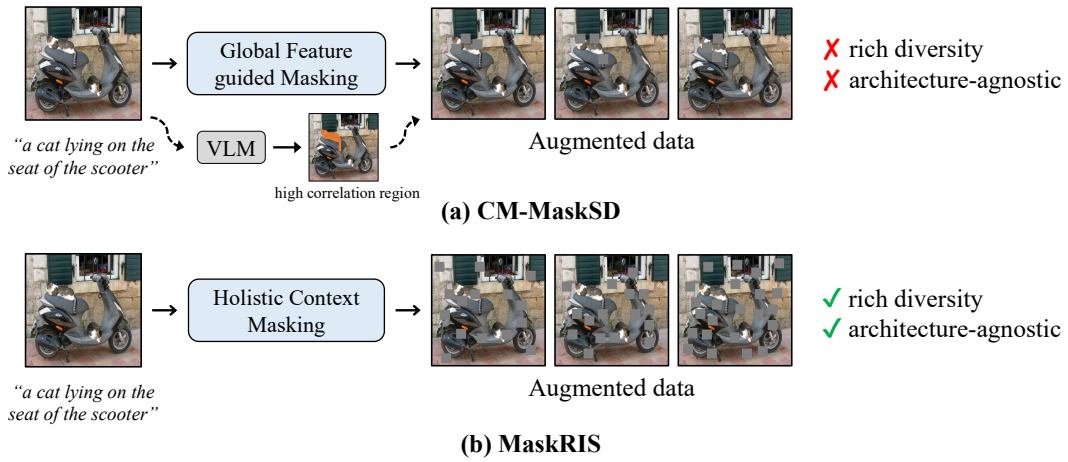


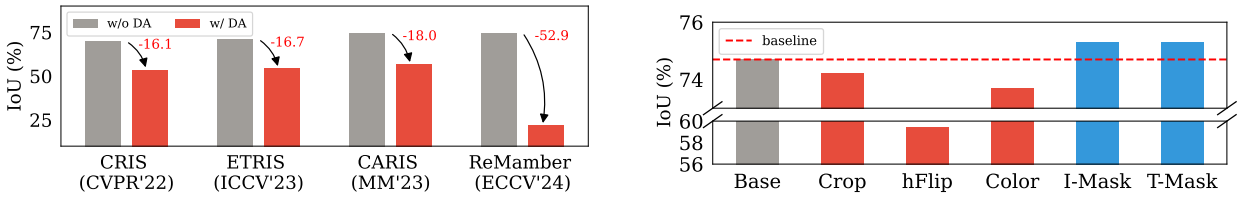
Figure 2: Comparison with CM-MaskSD (Wang et al., 2024). (a) CM-MaskSD employs global feature guided masking via a vision-language model (VLM), such as CLIP (Radford et al., 2021), which requires a pre-aligned vision–language representation. It focuses on high-correlation regions but lacks diversity and architecture-agnostic properties. (b) MaskRIS adopts holistic context masking, yielding richer diversity and improved architecture independence.

MaskSD employs masking guided by cross-modal correlations, where a global text feature is compared against all image patch tokens (and vice versa for a global image feature with text tokens). This correlation-based masking selects the most text-relevant patches and distills them through the [CLS]-token representation, resulting in an object-centric, token-level self-distillation scheme. However, this design inherently relies on a pre-aligned vision–language representation, limiting its generalizability. In contrast, MaskRIS introduces holistic context masking across both image and text, which not only preserves spatial cues and avoids augmentation-induced semantic flips (*e.g.*, "left" and "right", color changes), but also yields richer diversity by exposing the model to a broader range of masked contexts. Our Distortion-aware Contextual Learning (DCL) aligns *segmentation outputs* from a clean primary path and a masked secondary path, enforcing pixel-wise consistency while maintaining stability and diversity. This enables reasoning under incomplete or ambiguous multi-modal cues; empirically, masking both modalities outperforms masking only one (see ablations). Importantly, MaskRIS is free from pre-aligned cross-modal features, architecture-agnostic, and adds no parameters or inference overhead, thereby departing clearly from prior masking-based RIS methods.

Our contributions are as follows: (1) We provide a comprehensive analysis of data augmentation techniques for robust RIS training, identifying limitations in previous methods for the first time. (2) We propose a holistic context masking strategy that perturbs both image and text, increasing data diversity through varied, yet coherent partial contexts while reducing semantic conflicts across modalities. (3) We introduce Distortion-aware Contextual Learning (DCL), a dual-path distillation framework that enforces pixel-level consistency between original and masked predictions. Unlike prior [CLS] token–dependent schemes tied to specific architectures, our approach combines holistic masking with DCL to enable stable and architecture-agnostic training. (4) Our method achieves significant performance gains over existing RIS methods, including weakly supervised approaches, establishing new state-of-the-art results on RefCOCO, RefCOCO+, and RefCOCOg datasets. Moreover, we show that the proposed framework also benefits Referring Expression Comprehension (REC), highlighting its broader applicability beyond RIS.

2 Related Work

Referring Image Segmentation. Referring Image Segmentation (RIS) aims to segment objects in images based on text descriptions. This field evolved from early CNN-RNN/LSTM combinations for image and text processing (Margffoy-Tuay et al., 2018; Shi et al., 2018; Hu et al., 2016; Li et al., 2018) to advanced approaches using cross-modal attention (Liu et al., 2017; 2021a; Ye et al., 2019; Hu et al., 2020; Hui et al., 2020; Ding et al., 2021) and pre-trained transformer encoders (Wang et al., 2022; Kim et al., 2022; Yang



(a) Performance w/ conventional augmentations

(b) Performance w/ individual augmentation on CARIS

Figure 3: Existing RIS methods show a noticeable decline in their performance when applying conventional image augmentations (random cropping, color jittering, and horizontal flipping). In contrast, image masking (I-Mask) and text masking (T-Mask) improve model performance.

et al., 2022; Liu et al., 2023c). While previous studies have focused on image-text alignment, our study shifts focus to the training strategy in RIS, offering a complementary approach to enhance existing studies. This orthogonal perspective allows us to enjoy the performance benefits established by previous methods. Our concurrent work, NeMo (Ha et al., 2024), introduces a mosaic-based augmentation that combines a target image with carefully selected negative samples. While effective, it requires additional overhead for constructing and retrieving from a negative sample pool. In contrast, our approach provides a more efficient framework that enhances data diversity with a broader analysis of augmentation types and model robustness.

Masked Input Training. Masked input training is a self-supervised learning technique where random parts of input data are masked, and the model learns to predict these masked elements from visible data. Originally successful in natural language processing (NLP) as masked language modeling (MLM) (Devlin et al., 2018; Liu et al., 2019), it expanded to computer vision as masked image modeling (MIM) showing both training efficiency and strong generalization performance (Bao et al., 2021; Xie et al., 2022; He et al., 2022). These masking strategies are also employed to develop robust feature representations. RandomErasing (Zhong et al., 2020) and Cutout (DeVries & Taylor, 2017) erase randomly selected rectangle areas to reduce overfitting and handle occlusion. MIC (Hoyer et al., 2023) leverages context clues from the target domain through masked image consistency for domain adaptation. SMKD (Lin et al., 2023) uses masked knowledge distillation for few-shot classification.

Recently, MagNet (Chng et al., 2024) applies masked text training to improve vision-language alignment in RIS, where masked words are predicted using image features. Similarly, CM-MaskSD (Wang et al., 2024) applies masked input training using object-centric masking and CLIP-based patch-text alignment to refine feature representations. Although both methods improve cross-modal learning, their main focus is on feature alignment. In contrast, MaskRIS is inspired by the semantic conflicts from data augmentation in RIS. To address this issue, we performed a systematic analysis of augmentation-induced inconsistencies and their impact on RIS performance. Based on these findings, MaskRIS introduces a holistic masking strategy and a distortion-aware training framework, explicitly designed to improve model robustness in RIS.

Self-distillation. Self-distillation leverages the model itself as both teacher and student, preventing the need for separate teacher networks to facilitate knowledge distillation (Hinton et al., 2015). PS-KD (Kim et al., 2021) demonstrates the utility of softening hard targets through teacher model predictions, acting as an effective regularization strategy. Some studies (Zhang et al., 2019; Phuong & Lampert, 2019) explored self-distillation by employing an entire network as the teacher and an early exit network as the student. CoSub (Touvron et al., 2023) improved model performance in visual recognition tasks with sub-model-based self-distillation. MaskSub (Heo et al., 2023) introduced a drop-based technique for sub-model self-distillation, improving model performance and cost efficiency for image classification tasks. To fully exploit the benefits of input masking, we employ self-distillation between the predictions of the original and masked input.

3 Method

In this section, we propose a novel training strategy for RIS, called Masked Referring Image Segmentation (MaskRIS). MaskRIS aims to address the limitation of previous RIS (Sec. 3.1) with two key components:

(1) input masking strategy (Sec. 3.2) for data augmentation in RIS and (2) a dual-path training (Sec. 3.3) to maximize the benefits of the masking strategy.

3.1 Referring Image Segmentation (RIS)

RIS Pipeline. Given an image $x \in \mathbb{R}^{H \times W \times 3}$ and a text description that specifies an object within the image, the RIS model outputs a pixel-level mask delineating the referred object. The text description is first tokenized into a sequence of words $w = \{w_1, w_2, \ldots, w_L\}$, where $w_i \in \mathbb{R}^D$ represents a word embedding with D as the embedding dimension and L as the number of words in the description. Here, H and W denote the height and width of the image, and 3 represents the RGB color channels. We formalize this process as follows:

$$\hat{y} = f_\theta(x, w), \tag{1}$$

where $\hat{y} \in \mathbb{R}^{H \times W}$ is the predicted mask, generated through the interaction between the input image x and the tokenized words w by the neural network f parameterized by θ.

In conventional RIS approaches, f consists of the image encoder (Liu et al., 2021b; Radford et al., 2021) and the text encoder (Devlin et al., 2018; Radford et al., 2021) to extract high-dimensional visual and language features, along with a vision-language fusion module that integrates these features to generate the segmentation mask. Training the RIS models is essentially a pixel-wise binary classification task using the cross-entropy loss function:

$$\mathcal{L}_{ce}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{2}$$

where N denotes the total number of pixels, y_i is the binary label at each pixel i in the ground truth mask (1 indicating the object, 0 the background), and $\hat{y}_i$ is the predicted probability of the pixel belonging to the object.

Limited Data Augmentation in RIS. Previous studies (Yang et al., 2022; Wang et al., 2022; Liu et al., 2023c) have only focused on aligning vision and language features to bridge the modality gap. Still, effective training techniques for RIS remain underexplored, limiting the potential of vision-language features. As vision-language models grow in size and complexity (Radford et al., 2021; Jia et al., 2021; Alayrac et al., 2022), optimizing these training techniques becomes more crucial. Among training techniques, we regard data augmentation as the most promising approach, with the potential to enhance model robustness and generalization (Rebuffi et al., 2021; Li & Spratling, 2023; Yun et al., 2019; Cubuk et al., 2018; 2020).

Current RIS methods (Wang et al., 2022; Xu et al., 2023; Yang et al., 2022; Liu et al., 2023c) employ only simple resizing operations for data augmentations and don't utilize complex augmentations used in semantic segmentation. This is due to the fact that the complex data augmentations in segmentation are incompatible with RIS as shown in Figure 1, often altering or eliminating the referred objects. Consequently, these augmentations lead to a decline in RIS model performance on RefCOCO as demonstrated in Figure 3. Thus, we believe tackling the limited data augmentation problem is essential for overcoming a major performance bottleneck in RIS.

3.2 Masking Strategy for Image and Text

To overcome the limitations, we introduce a masking-based augmentation that combines both image and text masking to generate diverse image-text training pairs. Unlike conventional data augmentation, which often distorts essential spatial and attribute details, image masking minimizes such distortions, preserving this critical information while significantly expanding data diversity. In addition, image masking is beneficial in recognizing partially visible or occluded objects, as illustrated in Figure 4(a) and (b). State-of-the-art models, such as CARIS (Liu et al., 2023c), often struggle to identify partially visible objects. This issue highlights a prevalent challenge: the presence of occlusions often leads to the failure of models trained on

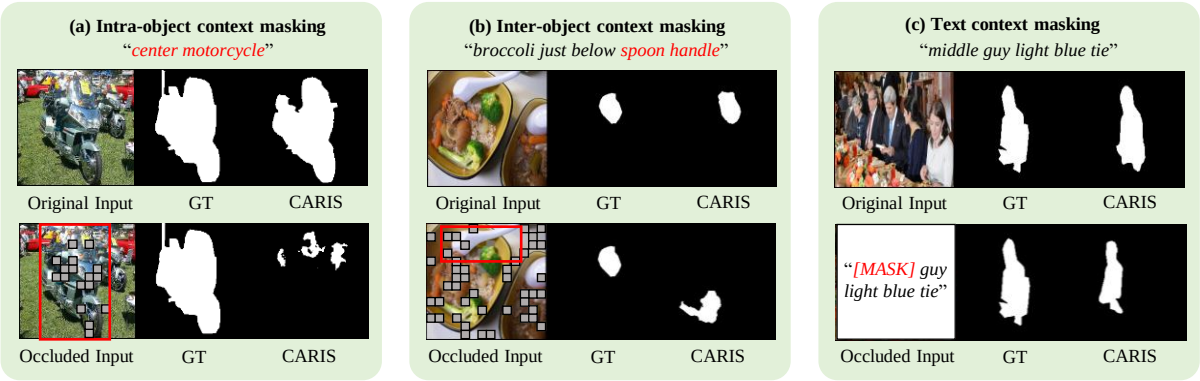


Figure 4: The existing RIS method tends to be inaccurate when faced with occluded context. CARIS (Liu et al., 2023c) represents the SoTA method in RIS. Words highlighted in red represent occluded objects in the image (left and center) and masked words in the text query (right).

clear and unoccluded objects. Similarly, as seen from Figure 4(b), text descriptions often rely on contextual objects (*e.g.*, "*spoon handle*" to describe the position of "*broccoli*"), suggesting that the impact of occlusion extends beyond the target objects themselves.

Beyond managing image occlusions, RIS requires flexibility in interpreting varied text descriptions, as a single referred object can be described in multiple ways. This diversity in textual description is essential for generalizing to different referring expressions. While our image masking addresses visual occlusions, we introduce text masking to enhance the model's capability to interpret diverse and incomplete descriptions by training it to rely on broader contextual cues. By masking parts of the text, we encourage the model to infer missing details, improving its understanding and reducing dependence on specific terms. For instance, as shown in Figure 4(c), CARIS can accurately identify a "*middle guy light blue tie*" with complete text. However, it struggles with partial descriptions, such as "[MASK] *guy light blue tie*", indicating the limitations in handling incomplete textual clues. By incorporating text masking, we enable the model to adapt to various referring expressions, enhancing its comprehension and robustness in RIS tasks.

We next describe our image and text masking techniques in detail, demonstrating how they address the aforementioned challenges in RIS. In the following, we specify the detailed process for both image and text masking, clarifying how these techniques are implemented to overcome the identified challenges and enhance the RIS model's performance.

Image Masking. We divide the input image x into non-overlapping patches and randomly mask a fixed ratio of patches. Following the MAE sampling strategy (He et al., 2022), we create a binary mask $M \in \mathbb{R}^{H \times W}$ by sampling patches uniformly without replacement, where 1 indicates a masked patch. The masked image x^M is obtained by element-wise multiplication: $x^M = (1 - M) \odot x$, where $\odot$ denotes the element-wise multiplication operation. This mechanism ensures that only a subset of the image's patches is exposed to the model during training, compelling the neural network to infer missing information and thereby enhancing its robustness and predictive accuracy.

While there might be concerns that our approach could hinder training by completely masking target objects, such cases are extremely rare in practice. For example, even with a relatively high masking ratio of 75%, the probability of a target object being completely masked is only 0.19% - remarkably low. This probability becomes even lower with reduced masking ratios. Furthermore, this potential issue is mitigated during training as the same samples are seen multiple times. These statistics support that our masking approach effectively augments the training data while maintaining crucial visual information.

Text Masking. We mask word tokens to generate masked word tokens $w^M = \{w_1^M, w_2^M, \ldots, w_L^M\}$ using a probabilistic strategy in MLM (Devlin et al., 2018). Specifically, each word w_i has a 15% chance of being masked. Of these, 80% are replaced with a [MASK] token, making it invisible to the model during training; 10% are replaced with a random word from the vocabulary, introducing noise and variability; and 10%

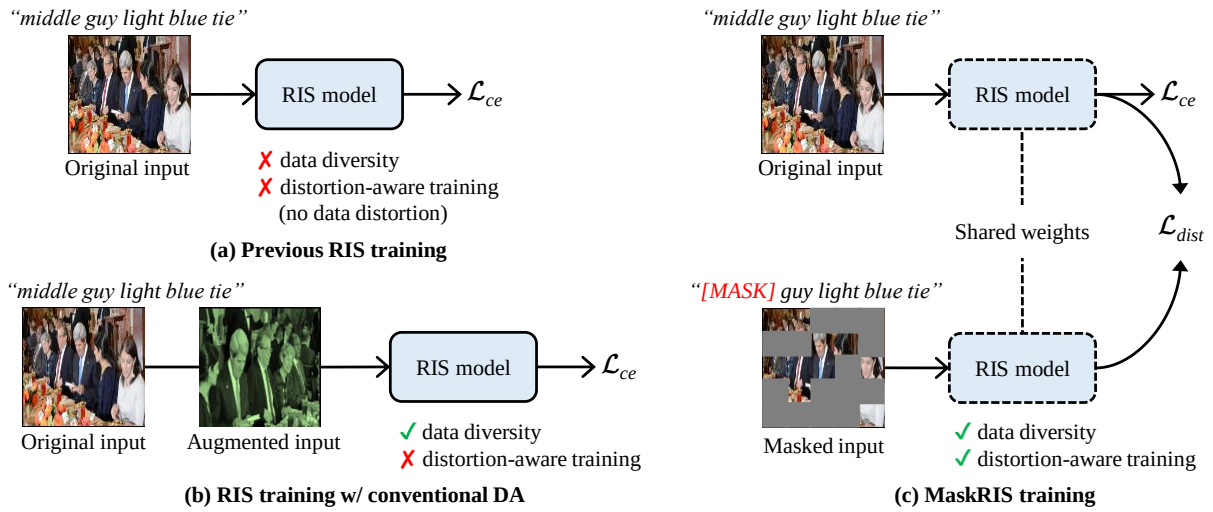


Figure 5: The overall framework of MaskRIS. Both image and text masking are employed to generate diverse image-text training pairs (Sec. 3.2). To maximize the benefits of the masking strategy, Distortion-aware Contextual Learning (DCL) is introduced (Sec. 3.3).

remain unchanged, preserving the original context. This strategy allows the model to experience a variety of textual scenarios, enhancing its ability to comprehend context and meaning even when parts of the text are masked or altered. Notably, as Wei & Zou (2019) witnessed, random word masking effectively preserves sentence meaning, further validating our method's balance between sentence coherence and data diversity.

3.3 Distortion-aware Contextual Learning

In the training phase, we employ both image and text masking simultaneously, enriching the diversity of the training dataset. Specifically, we formalize the processing of masked inputs through the model as follows:

$$\hat{y}^M = f_\theta(x^M, w^M), \tag{3}$$

where $\hat{y}^M$ denotes the model's output for the masked inputs. By adopting this strategy, MaskRIS learns to effectively segment objects under varying conditions of occlusion and incomplete descriptions. This simple integration effectively mitigates the intrinsic challenge of RIS that stems from the absence of an effective data augmentation strategy.

To further facilitate the RIS training, we introduce Distortion-aware Contextual Learning (DCL), as shown in Figure 5. This approach uses a primary path for processing original inputs and a secondary path for masked inputs. The primary path focuses on preserving the original image context, ensuring training stability, and promoting distortion-aware training. Meanwhile, the secondary path introduces variability by processing masked data, which improves the model's robustness. The training objective of MaskRIS consists of two parts, the distillation loss $\mathcal{L}_{dist}$ and the original pixel-wise classification loss $\mathcal{L}_{ce}$ in Eq. 2, *i.e.*, $\mathcal{L}_{total} = \mathcal{L}_{dist} + \mathcal{L}_{ce}$. In practice, we adopt a weighting factor between $\mathcal{L}_{ce}$ and $\mathcal{L}_{dist}$. While our default setting uses $\lambda = 0.5$, we also provide a detailed sensitivity analysis of this ratio in Appendix A.5, showing that MaskRIS is robust across a wide range of choices.

The distillation loss is defined through the binary cross-entropy loss between the predictions from both original and masked inputs as follows:

$$\mathcal{L}_{dist} = \mathcal{L}_{ce}(\mathsf{sg}(\hat{y}), \hat{y}^M) = -\frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \log(\hat{y}_i^M), \tag{4}$$

Method	Image Encoder	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
Standard: Training on the training split of each dataset.										
mIoU										
CRIS (Wang et al., 2022)	RN101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
ETRIS (Xu et al., 2023)	RN101	CLIP	71.06	74.11	66.66	62.23	68.51	52.79	60.28	60.42
RefTR (Li & Sigal, 2021)	RN101	BERT	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39
LAVT (Yang et al., 2022)	Swin-B	BERT	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
CM-MaskSD (Wang et al., 2024)	ViT-L	CLIP	74.89	77.54	71.28	67.47	71.80	59.91	66.53	66.63
CGFormer (Tang et al., 2023)	Swin-B	BERT	76.93	78.70	73.32	68.56	73.76	61.72	67.57	67.83
MaskRIS	Swin-B	BERT	**78.35**	**80.24**	**76.06**	**71.68**	**76.73**	**64.50**	**69.31**	**69.42**
oIoU										
LAVT (Yang et al., 2022)	Swin-B	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
CGFormer (Tang et al., 2023)	Swin-B	BERT	74.75	77.30	70.64	64.54	71.00	57.14	64.68	65.09
LQMFormer (Shah et al., 2024)	Swin-B	BERT	74.16	76.82	71.04	65.91	71.84	57.59	64.73	66.04
NeMo (Ha et al., 2024)	Swin-B	BERT	74.48	76.32	71.51	62.86	69.92	55.56	64.40	64.80
ReMamber (Yang et al., 2024)	Mamba-B	CLIP	74.54	76.74	70.89	65.00	70.78	57.53	63.90	64.00
CARIS† (Liu et al., 2023c)	Swin-B	BERT	74.67	77.63	71.71	66.17	71.70	57.46	64.66	65.45
MagNet (Chng et al., 2024)	Swin-B	BERT	75.24	78.24	71.05	66.16	71.32	58.14	65.36	66.03
MaskRIS	Swin-B	BERT	**76.49**	**78.96**	**73.96**	**67.54**	**74.46**	**59.39**	**65.55**	**66.50**
Combined: Training on the combination of three datasets.										
oIoU										
PolyFormer (Liu et al., 2023b)	Swin-B	BERT	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
CARIS† (Liu et al., 2023c)	Swin-B	BERT	76.67	79.85	73.07	68.23	73.98	59.59	67.42	68.73
MaskRIS	Swin-B	BERT	**78.71**	**80.64**	**75.10**	**70.26**	**75.15**	**62.83**	**69.12**	**71.09**

Table 1: Comparison with state-of-the-art methods on three benchmark datasets. † denotes the reproduced results across all experiments.

where sg is the stop-gradient function, which blocks gradient back-propagation. Inspired by Heo et al. (2023), MaskRIS employs a self-distillation framework that leverages soft targets from the original inputs, reducing training complexity and enhancing model generalization. This approach allows MaskRIS to effectively mitigate the learning instability caused by input masking while maintaining the benefits of increased data diversity and robustness.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate our method using three popular benchmarks in RIS: RefCOCO, RefCOCO+, and RefCOCOg. RefCOCO (Yu et al., 2016) provides 142,209 expressions (3.61 words per expression) for about 50,000 objects across 19,994 images, emphasizing object locations and appearances. RefCOCO+ (Yu et al., 2016) offers over 141,564 expressions (3.53 words per expression), excluding spatial language to focus on object attributes. RefCOCOg (Yu et al., 2016) introduces more complexity into the evaluation by including 85,474 longer and more complex expressions (8.43 words per expression) for 54,822 objects in 26,711 images. All datasets are collected on top of the MS COCO dataset (Lin et al., 2014). For RefCOCOg, we report results on the UMD partition (Yu et al., 2016), following the previous studies (Wang et al., 2022; Liu et al., 2023b; Kim et al., 2023b).

Evaluation Metrics. To evaluate model performance, we use mean Intersection-over-Union (mIoU), overall Intersection-over-Union (oIoU), and P@X. While mIoU calculates the average IoU across all test samples, oIoU measures the ratio of the total intersection to the total union areas across all samples. P@X measures the percentage of test samples with an IoU above a threshold $\mathsf{X} \in \{0.5, 0.7, 0.9\}$. These metrics ensure a consistent and fair evaluation across different RIS methods.

Implementation Details. To validate MaskRIS, we applied it to various RIS models (Xu et al., 2023; Wang et al., 2022; Yang et al., 2022; Liu et al., 2023c). Designed as a plug-and-play training strategy, we strictly follow the original training settings and hyperparameters, such as learning rate, epochs, and batch size, without modification. Notably, we primarily implemented our method on CARIS (Liu et al., 2023c),

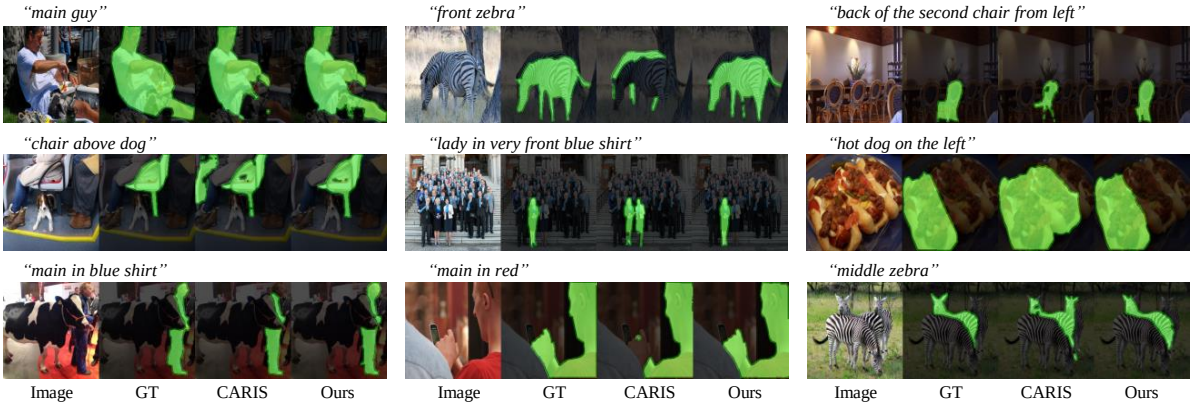


Figure 6: Qualitative examples of segmentation results on the RefCOCO dataset.

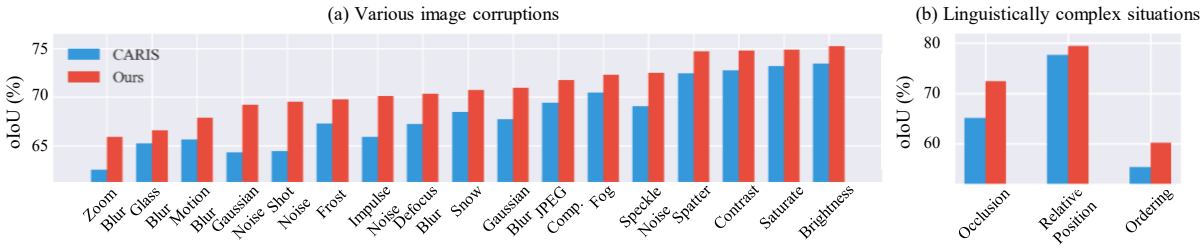


Figure 7: Robustness on various image corruptions provided by ImageNet-C (Hendrycks & Dietterich, 2019) and linguistically complex situations. The results (oIoU) are evaluated on the RefCOCO validation set.

a leading SoTA method, unless stated otherwise. Images are resized to 448 × 448 for both training and testing. For image masking, we set 32 as the patch size.

4.2 Main Results

Comparison with State of the Arts. We compare MaskRIS with previous methods on three popular benchmarks. As shown in Table 1, MaskRIS consistently outperforms previous methods by significant margins. Specifically, compared to the second-best performing method, CARIS (Liu et al., 2023c), our approach improves oIoU scores by 1.82%p, 1.33%p, and 2.25%p on the RefCOCO validation, testA, and testB sets, respectively. On RefCOCO+, our method leads by 1.37%p, 2.76%p, and 1.93%p on the validation, testA, and testB splits, respectively. Even on the challenging RefCOCOg dataset, MaskRIS still outperforms CARIS by 0.89%p and 1.05%p on the validation and test splits. These results validate the effectiveness of our masking strategy for RIS model training.

To further demonstrate the capability of our method, we conduct experiments on a larger, more comprehensive dataset. This dataset is a combination of the training sets from RefCOCO, RefCOCO+, and RefCOCOg, following the approach used in previous studies (Liu et al., 2023b;c). Training on this combined dataset yields even better results, with oIoU improvements of 1.17–4.04%p over previous methods. This result indicates that our method is effective across different types of data and highlights its generalizability. Figure 6 shows qualitative examples on RefCOCO, where our method more comprehensively captures the extent of objects by expanding or reducing their coverage (1st row and 2nd row). Additionally, it demonstrates robustness against occlusion (3rd row) and precisely identifies target objects in alignment with the provided text descriptions (4th row). These results confirm the ability of our approach to enhance model robustness and leverage textual cues effectively.

Robustness Evaluation. A key strength of our approach lies in its robustness across visually and linguistically complex scenarios. To this end, we first evaluate our model's robustness to various image corruptions. We use the benchmark provided by ImageNet-C (Hendrycks & Dietterich, 2019), which includes 17 types of corruption, each with 5 severity levels. To evaluate the performance, we compute the oIoU at each severity

Method	Image Encoder	Text Encoder	RefCOCO		
			val	testA	testB
Weakly-supervised approach					
mIoU					
TRIS (Liu et al., 2023a)	RN50	CLIP	31.17	32.43	29.56
TRIS (Liu et al., 2023a)+MaskRIS	RN50	CLIP	32.60 (+1.43)	34.26 (+1.83)	30.32 (+0.76)
SaG (Kim et al., 2023a)	ViT-S/16	BERT	37.21	36.60	39.41
SaG (Kim et al., 2023a)+MaskRIS	ViT-S/16	BERT	38.85 (+1.64)	37.70 (+1.1)	40.20 (+0.79)
Fully-supervised approach					
mIoU					
CRIS (Wang et al., 2022)	RN50	CLIP	69.52	72.72	64.70
CRIS (Wang et al., 2022)+MaskRIS	RN50	CLIP	70.73 (+1.21)	74.06 (+1.34)	66.82 (+2.12)
ETRIS (Xu et al., 2023)	R101	CLIP	71.06	74.11	66.66
ETRIS (Xu et al., 2023)+MaskRIS	R101	CLIP	72.39 (+1.33)	74.99 (+0.88)	68.55 (+1.89)
oIoU					
LAVT (Yang et al., 2022)	Swin-B	BERT	72.73	75.82	68.79
LAVT (Yang et al., 2022)+MaskRIS	Swin-B	BERT	73.79 (+1.06)	76.57 (+0.75)	70.31 (+1.52)
ReMamber† (Yang et al., 2024)	Mamba-B	CLIP	72.37	74.85	68.29
ReMamber (Yang et al., 2024)+MaskRIS	Mamba-B	CLIP	74.02 (+1.65)	76.32 (+1.47)	69.06 (+0.77)
DETRIS† (Huang et al., 2025)	DINOv2-B	CLIP	74.47	77.40	71.46
DETRIS (Huang et al., 2025)+MaskRIS	DINOv2-B	CLIP	76.08 (+1.61)	78.46 (+1.06)	72.84 (+1.38)
CARIS† (Liu et al., 2023c)	Swin-B	BERT	74.67	77.63	71.71
CARIS (Liu et al., 2023c)+MaskRIS	Swin-B	BERT	76.49 (+1.82)	78.96 (+1.33)	73.96 (+2.25)

Table 2: Compatibility of MaskRIS with various RIS methods. MaskRIS consistently enhances existing methods in both fully supervised and weakly supervised settings.

level and average these scores for each corruption type. As shown in Figure 7(a), MaskRIS consistently outperforms the previous SoTA method, CARIS (Liu et al., 2023c), with oIoU improvements of 1.34–5.07%p. This demonstrates that MaskRIS effectively reduces overfitting to clean data, providing greater robustness against a wide range of visual distortions.

In addition, we evaluate MaskRIS's robustness to linguistically complex situations, such as occlusion, relative position, and ordering. For occlusion scenarios, we modify the input data by occluding parts of objects. For relative position and ordering, we select samples containing relevant linguistic keywords from the validation set. As shown in Figure 7(b), MaskRIS achieves superior performance in these complex scenarios on the RefCOCO validation set, further highlighting its robustness to both visual and linguistic complexity.

Compatibility with Other Methods. Our training strategy is highly compatible, integrates seamlessly with various RIS methods, and achieves significant performance improvements across both weakly supervised and fully supervised settings, as shown in Table 2. In the weakly supervised regime, where models rely solely on textual descriptions without ground-truth masks, MaskRIS still yields notable gains. For instance, TRIS (Liu et al., 2023a) improves by +1.43%p on the validation set, +1.83%p on testA, and +0.76%p on testB. Likewise, SaG (Kim et al., 2023a) is enhanced by up to +1.64%p, reaching 38.85 mIoU on the validation set. In the fully supervised setting, MaskRIS demonstrates equally strong compatibility. CRIS (Wang et al., 2022) shows improvements of +1.21%p (val), +1.34%p (testA), and +2.12%p (testB), while ETRIS (Xu et al., 2023) gains up to +1.89%p. Beyond CNN-based baselines, transformer-based approaches such as LAVT (Yang et al., 2022) and CARIS (Liu et al., 2023c) also benefit significantly, with CARIS achieving the highest improvements of +2.25%p on testB. More recent architectures, including ReMamber (Yang et al., 2024) and DETRIS (Huang et al., 2025), further validate the generality of our method, consistently obtaining +0.8–2.0%p gains without architectural modifications. Overall, MaskRIS consistently improves RIS model performance across diverse architectures and training paradigms, demonstrating its adaptability and effectiveness without modifying the original designs.

Compatibility with REC Methods. The conflict between conventional image augmentations and language-grounded tasks is not unique to RIS. Referring Expression Comprehension (REC) also relies on

Method	Image Encoder	Text Encoder	RefCOCO		
			val	testA	testB
DetAcc					
EEVG† (Chen et al., 2024)	ViT-B	BERT	87.87	89.97	85.34
EEVG (Chen et al., 2024)+MaskRIS	ViT-B	BERT	88.72 (+0.85)	90.81 (+0.84)	85.36 (+0.02)
SimVG† (Dai et al., 2024)	ViT-B	BEiT-3	86.94	88.97	82.87
SimVG (Dai et al., 2024)+MaskRIS	ViT-B	BEiT-3	87.79 (+0.85)	90.02 (+1.05)	84.88 (+2.01)
C3VG† (Dai et al., 2025)	ViT-B	BEiT-3	92.42	94.12	89.34
C3VG (Dai et al., 2025)+MaskRIS	ViT-B	BEiT-3	93.29 (+0.87)	94.86 (+0.74)	90.29 (+0.95)

Table 3: Compatibility of MaskRIS with REC methods. MaskRIS consistently enhances existing methods.

Figure 8: Qualitative examples of masking strategies on the RefCOCO dataset. I-Mask, T-Mask, and Both denote the results of applying image masking, text masking, and both, respectively.

precise spatial and attribute alignment between text and image. For example, flipping an image containing the phrase "the man on the left" alters the semantic meaning for both RIS and REC. This confirms that the augmentation dilemma is indeed universal to referring tasks. To investigate this, we integrated MaskRIS into various REC methods (Chen et al., 2024; Dai et al., 2024; 2025). As shown in Table 3, MaskRIS consistently improves performance when combined with existing REC models by enhancing robustness to occlusion and incomplete linguistic cues while increasing data diversity. The findings confirm that MaskRIS is not task-specific. Its masking approach addresses the augmentation dilemma across both RIS and REC, making it a general training strategy for referring tasks where traditional augmentations are unsafe.

4.3 Ablation Study

Impact of Each Component of MaskRIS. We analyze the impact of each MaskRIS component in model training. Table 4 shows segmentation results on the RefCOCO validation set. The results clearly show that masking images and text separately improves model performance. However, the most significant improvement occurs when we apply masking to both images and text together. This demonstrates that the combination of visual and textual masking strategies significantly improves the model's ability to accurately identify and segment the referred objects. In addition, Distortion-aware Contextual Learning (DCL) provides further performance gains. It effectively mitigates the learning instability caused by input masking, while maintaining the benefits of increased data diversity and robustness.

Figure 8 illustrates the benefits of image and text masking. CARIS often fails to correctly identify objects when they or their adjacent objects are partially obscured. However, by incorporating image masking, our approach achieves accurate object recognition (1st and 2nd rows). For example, in the scenario shown in the first row with a *"baby girl"* whose legs are obscured, CARIS inaccurately identifies the legs of an adult as part of the target object. On the other hand, our method successfully separates the obscured elements, demonstrating the superior accuracy of our method in complex environments. Conversely, text

IM	TM	DCL	P@0.5	P@0.7	P@0.9	oIoU
✘	✘	✘	87.73	80.20	39.60	74.67
✔	✘	✘	87.76	80.52	38.73	75.31
✘	✔	✘	87.72	80.45	39.26	75.32
✔	✔	✘	88.00	81.35	40.11	75.71
✔	✘	✔	88.60	81.86	41.08	76.02
✘	✔	✔	88.34	81.19	39.24	75.76
✔	✔	✔	**88.62**	**81.95**	**41.19**	**76.49**

Table 4: Impact of each component of MaskRIS on the RefCOCO validation set, where IM (TM) refers to image (text) masking.

Masking Approach	Image Encoder	Text Encoder	oIoU
Object-centric masking	Swin-B	BERT	75.14
Context masking			75.93
Holistic masking			**76.02**

Table 5: Comparison of different image masking approaches on the RefCOCO validation set. Context masking refers to the random masking of non-target areas.

masking improves the alignment between the text description and the target object (3rd row). For example, with the *"left marine"*, CARIS inaccurately identifies the target object by focusing too much on spatial information (*i.e.*, *"left"*), as shown in Figure 4(c). Text masking intervenes in such cases, ensuring the accurate recognition of the target object. Furthermore, by integrating both image and text masking, our approach exploits the strengths of each masking strategy.

4.4 Discussion

Comparison with CM-MaskSD. MaskRIS differs from CM-MaskSD (Wang et al., 2024) in three key aspects: purpose, masking strategy, and adaptability. First, MaskRIS explicitly handles the semantic conflicts caused by conventional augmentations in RIS, ensuring that transformations do not alter the contextual meaning of referring expressions. In contrast, CM-MaskSD focuses on improving patch-text alignment using CLIP's global vector but does not directly tackle augmentation-induced inconsistencies. This makes semantic conflict resolution a unique contribution of MaskRIS. Second, while CM-MaskSD employs object-centric masking, selectively targeting key objects or keywords, MaskRIS utilizes a holistic masking approach that applies masking across all image regions. This strategy enhances model generalization and robustness. As shown in Table 5, holistic masking consistently outperforms object-centric masking. Third, CM-MaskSD is architecture-dependent due to its reliance on the [CLS] token, limiting its applicability to CLIP-based models. In contrast, MaskRIS is architecture-agnostic and can be seamlessly integrated into various RIS and REC frameworks.

5 Conclusion

In this study, we introduce Masked Referring Image Segmentation (MaskRIS), an effective training framework for Referring Image Segmentation (RIS) that combines image and text masking with Distortion-aware Contextual Learning (DCL). MaskRIS addresses the challenges of conventional data augmentation in RIS by minimizing semantic distortion and enhancing data diversity. This approach strengthens the model's robustness to occlusions and incomplete information, achieving new state-of-the-art accuracies on the RefCOCO, RefCOCO+, and RefCOCOg datasets. Our results demonstrate MaskRIS's effectiveness and adaptability in both fully and weakly-supervised settings, highlighting its potential as a versatile and powerful baseline for advancing the field of RIS.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8721–8729, 2018.

Wei Chen, Long Chen, and Yu Wu. An efficient and effective transformer decoder-based framework for multi-task visual grounding. In *European Conference on Computer Vision*, pp. 125–141. Springer, 2024.

Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26573–26583, 2024.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in neural information processing systems*, 37:121670–121698, 2024.

Ming Dai, Jian Li, Jiedong Zhuang, Xian Zhang, and Wankou Yang. Multi-task visual grounding with coarse-to-fine consistency constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2618–2626, 2025.

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

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16321–16330, 2021.

Seongsu Ha, Chaeyun Kim, Donghwa Kim, Junho Lee, Sangho Lee, and Joonseok Lee. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Byeongho Heo, Taekyung Kim, Sangdoo Yun, and Dongyoon Han. Augmenting sub-model to improve main model. *arXiv preprint arXiv:2306.11339*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11721–11732, 2023.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 108–124. Springer, 2016.

Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4424–4433, 2020.

Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3653–3661, 2025.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 59–75. Springer, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. Shatter and gather: Learning referring image segmentation with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15547–15557, 2023a.

Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6567–6576, 2021.

Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18145–18154, 2022.

Seoyeon Kim, Minguk Kang, and Jaesik Park. Risclip: Referring image segmentation framework using clip. *arXiv preprint arXiv:2306.08498*, 2023b.

Lin Li and Michael Spratling. Data augmentation alone can improve adversarial training. *arXiv preprint arXiv:2301.09879*, 2023.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2018.

Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, and Shih-Fu Chang. Supervised masked knowledge distillation for few-shot transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19649–19659, 2023.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1271–1280, 2017.

Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22124–22134, 2023a.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023b.

Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (9):4761–4775, 2021a.

Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Caris: Context-aware referring image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 779–788, 2023c.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 630–645, 2018.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2085–2094, October 2021.

Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1355–1364, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pp. 492–504. PMLR, 2023.

Nisarg A Shah, Vibashan VS, and Vishal M Patel. Lqmformer: Language-aware query mask transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12903–12913, 2024.

Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–54, 2018.

Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23570–23580, 2023.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Matthieu Cord, Maxime Oquab, Piotr Bojanowski, Jakob Verbeek, and Hervé Jégou. Co-training 2l submodels for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11701–11710, 2023.

Wenxuan Wang, Xingjian He, Yisi Zhang, Longteng Guo, Jiachen Shen, Jiangyun Li, and Jing Liu. Cm-masksd: Cross-modality masked self-distillation for referring image segmentation. *IEEE Transactions on Multimedia*, 26:6906–6916, 2024.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2628–2638, 2023.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17503–17512, 2023.

Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. *arXiv preprint arXiv:2403.17839*, 2024.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2022.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10502–10511, 2019.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *Advances in Neural Information Processing Systems*, 35:14729–14742, 2022.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.