Mask Grounding for Referring Image Segmentation

Yong Xien Chng[1,2] Henry Zheng[1] Yizeng Han[1] Xuchong Qiu[2†] Gao Huang[1✉]
[1]Department of Automation, BNRist, Tsinghua University [2]Bosch Corporate Research

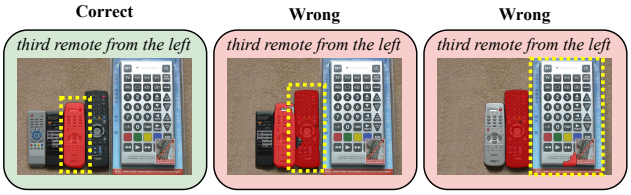
Abstract

Referring Image Segmentation (RIS) is a challenging task that requires an algorithm to segment objects referred by free-form language expressions. Despite significant progress in recent years, most state-of-the-art (SOTA) methods still suffer from considerable language-image modality gap at the pixel and word level. These methods generally 1) rely on sentence-level language features for language-image alignment and 2) lack explicit training supervision for fine-grained visual grounding. Consequently, they exhibit weak object-level correspondence between visual and language features. Without well-grounded features, prior methods struggle to understand complex expressions that require strong reasoning over relationships among multiple objects, especially when dealing with rarely used or ambiguous clauses. To tackle this challenge, we introduce a novel Mask Grounding auxiliary task that significantly improves visual grounding within language features, by explicitly teaching the model to learn fine-grained correspondence between masked textual tokens and their matching visual objects. Mask Grounding can be directly used on prior RIS methods and consistently bring improvements. Furthermore, to holistically address the modality gap, we also design a cross-modal alignment loss and an accompanying alignment module. These additions work synergistically with Mask Grounding. With all these techniques, our comprehensive approach culminates in MagNet (Mask-grounded Network), an architecture that significantly outperforms prior arts on three key benchmarks (RefCOCO, RefCOCO+ and G-Ref), demonstrating our method's effectiveness in addressing current limitations of RIS algorithms. Our code and pre-trained weights will be released.

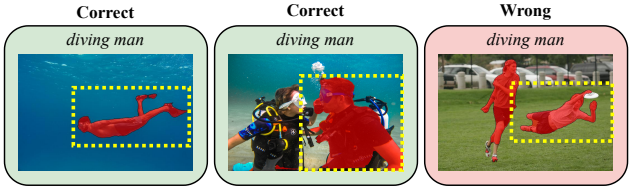
1. Introduction

Deep learning has greatly improved the performance of vision algorithms on many image segmentation tasks, such as semantic segmentation [5, 48], instance segmentation [2, 12, 24, 42] and panoptic segmentation [8, 36]. These tasks require grouping of image pixels under a fixed set of pre-defined categories and mainly differ in the granularity of grouping semantics required. In contrast to these uni-modal segmentation tasks, Referring Image Segmentation (RIS) [9, 28] is a challenging multi-modal task that requires an algorithm to simultaneously understand fine-grained human language expression and make correct pixel-level correspondence to the referred object. Recently, it has gained widespread research attention due to its potential to improve human-robot interaction [1], interactive image editing [43, 52] and advanced driver-assistance systems [29].

The key challenge in RIS lies in how to reduce the modality gap between language and image features [14, 64, 71]. To tackle this challenge, we need to have an effective alignment between a given language expression and the corresponding image pixels for highlighting the referred target. Ideally, with robust pixel-wise language-image alignment, language and image features should have high feature similarity when referring to the same object and low feature sim-



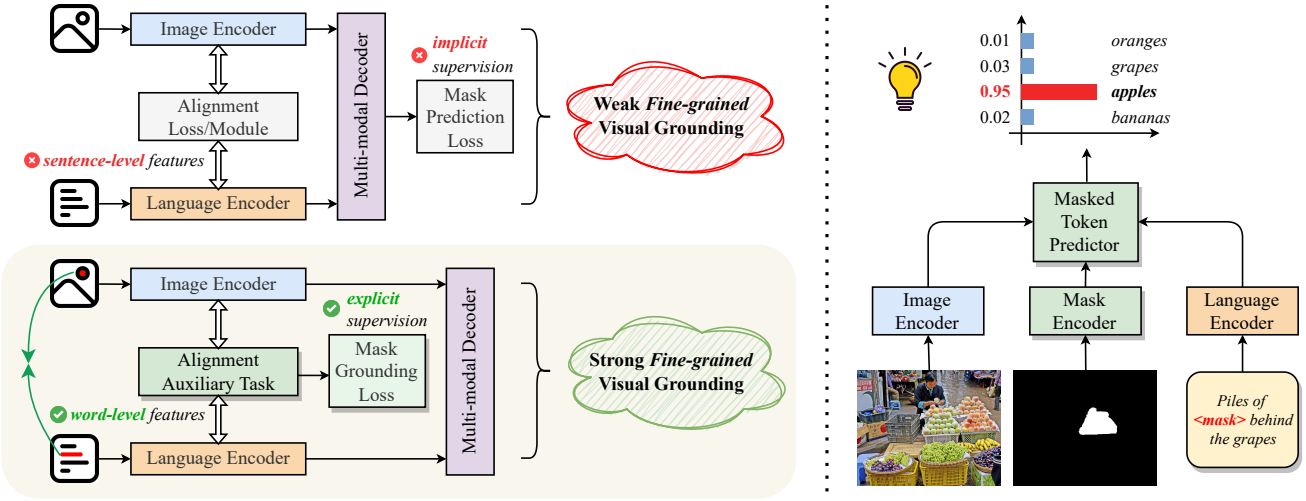
(a) *Fine-grained* visual grounding is required to reason over complicated relationships among multiple objects.



(b) *Fine-grained* visual grounding is required to understand expressions used in uncommon or ambiguous contexts.

Figure 1. Importance of Fine-grained Visual Grounding for RIS. Most RIS algorithms lack well-grounded text features. As a result, they struggle in difficult cases illustrated in (a) and (b). Red mask are predictions of LAVT, one of the recent SOTA RIS methods. Yellow dotted boxes are the ground truths.

† Project lead.
✉ Corresponding author.



(a) Comparison between our method and pervious methods

(b) Our proposed Mask Grounding

Figure 2. (a) Current SOTA RIS methods mainly focus on designing and improving multi-modal alignment modules and/or alignment losses. These methods generally 1) do not have explicit training supervision for fine-grained visual grounding and 2) use sentence-level language features or image/pixel-level image features for alignment. As a result, their language features lack precise visual-textual object correspondence. (b) Our proposed Mask Grounding remedies this problem by explicitly teaching our model to learn fine-grained correspondence between masked word tokens and their matching visual objects through an auxiliary alignment task.

ilarity when referring to different objects. However, achieving such alignment is non-trivial because the language expression can be highly complex and diverse.

As depicted in Fig. 2, prevailing methods primarily focus on devising innovative losses [61, 71] or introducing new network architectures/modules [14, 45, 63, 64, 72] to bolster language-image alignment. Despite their advancements, two overarching limitations persist. First, these approaches tend to rely on *sentence-level language features* for language-image alignment. Second, they often *lack explicit training supervision* for fine-grained visual grounding. These oversights result in their language features becoming noisy anchors for RIS prediction [63, 69], inhibiting the effective learning of fine-grained visual grounding. Consequently, such models face challenges when interpreting referring expressions that require intricate reasoning across complex inter-object relationships or contain clauses used in rare or ambiguous contexts, as exemplified in Fig. 1.

To address this challenge, we introduce a novel **Mask Grounding** auxiliary task to explicitly teach our model to make fine-grained correspondence between masked textual tokens and their matching visual objects. Specifically, during training, our model encounters randomly masked textual tokens and has to predict their identities. Instead of relying solely on the surrounding textual context to predict these tokens, our model integrates both visual and segmentation information. This integrated approach is pivotal for the model to make accurate prediction, as it must discern and establish the correct linkage between the masked tokens and their corresponding visual objects. Learning to do so ensures that our model acquires a profound proficiency in the highly-coveted fine-grained visual grounding. The efficacy of Mask Grounding is empirically validated with extensive ablation experiments. Moreover, we also show that Mask Grounding is universal and can be directly used on prior RIS methods to bring significant improvements.

In addition to Mask Grounding, we also design a cross-modal alignment loss and an alignment module to holistically bridge the modality gap. With all these enhancements, our resulting MagNet (Mask-grounded Network) sets new records by significantly outperforming previous SOTA methods across all key datasets (RefCOCO [67], RefCOCO+ [67] and G-Ref [49, 50]). Notably, our method consistently outperforms these SOTA methods by large margins of up to **2.48** points in overall IoU. Visual examination of MagNet's predictions reinforces our claim and shows that our method works well in complex scenarios.

Our main contributions are summarized as follows:

1. We highlight the shortcomings in recent state-of-the-art (SOTA) RIS algorithms, pinpointing the lack of fine-grained visual grounding.
2. We introduce the Mask Grounding auxiliary task, a novel method aimed at enhancing fine-grained visual grounding in existing RIS algorithms. Its effectiveness is validated through rigorous ablation studies.
3. Using Mask Grounding, together with our specially designed cross-modal alignment loss and an accompanying alignment module, we present MagNet (Mask-grounded Network), a new SOTA network for RIS.

2. Related works

Architecture Design for RIS. Early works [28, 41, 44, 54] follow a concatenate-then-convolve pipeline, where language and image features are fused by concatenation. Subsequent works [4, 41, 44, 65] improve upon this pipeline by using RNN or dynamic networks [19–21, 23] to progressively refine the segmentation mask. Other works [16, 64] investigate the position to perform language-image fusion and conclude that early fusion performs the best. Apart from designing novel fusion mechanisms, some works [31, 32, 68] exploit known linguistic structures or object relationships to enhance language-image fusion. Riding on the success of attention architecture [15, 22, 58], current works mostly use unidirectional [54, 64, 66] or bidirectional [30, 71] cross-attention modules to perform language-image fusion. To improve model performance on novel composition of learned concepts, a recent work [62] uses meta learning [17]. Driven by the success of large language models [3, 57], newer works [10, 46, 74] explore formulating RIS as an auto-regressive vertex generation problem. Lately, VPD [72] attempts to exploit semantic information in diffusion models [27, 51, 55] for RIS, whereas ReLA [45] and DMMI [29] generalize RIS to support an arbitrary number of targets. Despite huge progress in RIS architecture design, prior studies often expect language-image alignment to be performed implicitly through mask prediction. We enhance this by introducing an auxiliary task for explicit language-image feature alignment.

Loss Design for RIS. Early works train RIS models with simple binary cross entropy loss. Inspired by the success of prior works [59, 73] in adopting contrastive loss [6, 18, 25, 53] for semantic segmentation tasks, recent works [14, 71] start to use contrastive loss in order to regularize the segmentation embedding space and achieve good results. Contrary to prior works that use global-pooled language features for loss computation, we focus on learning fine-grained object correspondence at the pixel-word level.

Masked Language Modeling. Masked language modeling (MLM) is a powerful technique for natural language processing that trains a model to restore missing or corrupted tokens in an input text. It was introduced by BERT [13] and has become a popular technique for pre-training language [11, 70] and visual language [35, 39, 56] models. Recently, it has been shown to scale excellently [3, 33] and generalize well to various downstream tasks [3, 57]. A work closely related to ours is MaskedVLM [37], which is a multi-modal adaptation of MLM that jointly performs masked vision and language modeling. It does so by reconstructing the masked signal of one modality with the help from the another modality. Mask Grounding differs from MaskedVLM by using extra mask signals that directly match the missing words to ensure clear and meaningful reconstructions, so that fine-grained visual grounding can be effectively learnt.

3. Method

In this section, we first describe our architecture overview (Sec. 3.1). Then, we explain our proposed Mask Grounding (Sec. 3.2) auxiliary task, cross-modal alignment module (Sec. 3.3) and alignment loss (Sec. 3.4). Finally, we give the overall loss function Sec. 3.5 for our model.

3.1. Architecture Overview

MagNet (Mask-grounded Network) adopts a unified approach that integrates three inter-linked modules to enhance visual-textual object correspondence and segmentation accuracy. Mask Grounding is the first of these integrated modules, designed to improve fine-grained visual grounding in language features. It accomplishes this by teaching the model to predict masked textual tokens, using a combination of visual cues, linguistic context, and segmentation information. Building upon Mask Grounding's enriched language features, Cross-modal Alignment Module (CAM) steps in to fine-tune the bi-directional interaction between the refined language and image features. By incorporating global contextual information from multiple image scales, CAM ensures that the multi-modal features are in sync, addressing the granularity discrepancies between textual descriptions and visual information. Finally, Cross-modal Alignment Loss (CAL) cohesively weaves together pixel-to-pixel and pixel-to-text alignments. By simultaneously considering these alignments, CAL ensures that segments created by the model are not only accurate in shape but also correctly match their referring textual descriptions.

3.2. Mask Grounding

Inspired by prior works [13, 26, 60] that have shown the effectiveness of using masked input modeling to learn good feature representation, we propose a novel Mask Grounding auxiliary task to improve the learning of fine-grained visual grounding information in language features. As shown in Fig. 3, given an input image, its corresponding referring expression and segmentation mask, we randomly replace some tokens in the tokenized referring expression with a special learnable mask token and train our model to predict the actual tokens being masked. By successfully predicting the identities of masked tokens, our model will acquire the ability to understand which parts of the text correspond to which parts of the image, thus learning fine-grained visual grounding in the process. Specifically, to perform this auxiliary task, we first encode the segmentation mask into a mask embedding by first extracting the center coordinates of the mask region and passing it through a 2-layer MLP. At the same time, we use a linear layer to project the language embedding into the same dimension as the image embedding. Then, we employ the proposed Masked Token Predictor to jointly process all these concatenated embeddings with attention mechanism for masked token prediction. Finally, a

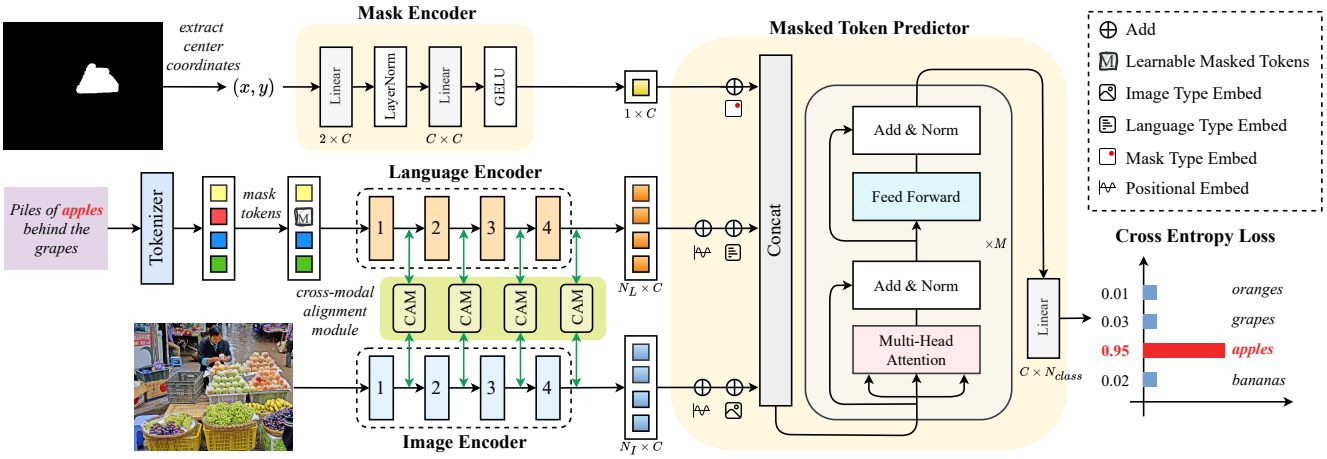


Figure 3. Overview of Mask Grounding. This task enriches fine-grained visual grounding in language features by guiding the model to learn detailed textual-visual associations. To perform this task, we first use an MLP-based Mask Encoder to encode center-coordinates of segmentation masks. Then, we randomly mask textual tokens in language inputs before extracting their features. Finally, we pass the encoded language, image and mask features to a Transformer-based Masked Token Predictor to perform masked token prediction.

cross-entropy loss $\mathcal{L}_{\text{grounding}}$ is used to compare the final predicted distribution with the target distribution. The large-scale BERT [13] vocabulary is adopted as our word class list, as it is generally accepted to have open-vocabulary capability. Although additional forward pass through the language encoder is required to process the masked expression, overall computational cost only increase by 4.76% as the language encoder is very small. We believe this slight increase in computational cost is an acceptable trade-off to improve visual grounding in language features.

A brief mathematical formulation for Mask Grounding can be given as follows: Let $\mathbf{T}$, $\mathbf{I}$, $\mathbf{M}$ be the input to the language encoder, image encoder and mask encoder,

$$\mathbf{O} = \text{LanguageEncoder}(\text{Mask}(\mathbf{T})), \quad (1)$$

$$\mathbf{P} = \text{ImageEncoder}(\mathbf{I}), \quad (2)$$

$$\mathbf{C} = \text{MaskEncoder}(\mathbf{M}), \quad (3)$$

$$\mathcal{L}_{\text{grounding}} = \mathcal{L}_{\text{CE}}(\mathbf{y}_{\text{gt}}, \text{Predictor}(\text{Concat}([\mathbf{O}, \mathbf{P}, \mathbf{C}])), \quad (4)$$

where Predictor is a BERT [13]-like encoder, M is the center coordinates of ground truth masks, $\mathbf{y}_{\text{gt}}$ is the label of the masked token and $\mathcal{L}_{\text{CE}}$ is the cross entropy loss. In our experiments, we use Swin-B [47] as our image encoder, and BERT-base [13] as our language encoder, but our approach is not specifically bound to these encoders.

Discussion. In Tab. 3(a), we demonstrate Mask Grounding's superiority over both the standard masked language modeling (MLM) [3, 13, 33, 57] and masked-vision language modeling (MaskedVLM) [37], highlighting our approach's effectiveness. Our advantages over these techniques include: 1) *Modality Integration*: Traditional MLM is uni-modal and lacks correspondence between referring expressions and their matching visual objects.. While MaskedVLM is multi-modal, Mask Grounding surpasses it by introducing an additional masking signal that aligns with the masked words and their matching visual objects, enabling a more coherent reconstruction. This approach exposes word-object correspondence and allows fine-grained visual grounding to be learnt. 2) *Task Nature*: MLM and MaskedVLM serve as general pre-training tasks and require fine-tuning for specific downstream applications. In contrast, Mask Grounding is designed as a specialized auxiliary task for RIS, enhancing fine-grained visual grounding within language features right from the training phase. Consequently, there is no need for additional fine-tuning. 3) *Prediction Context*: While MLM and MaskedVLM predict using textual or textual-visual contexts, Mask Grounding incorporates both with additional segmentation information. By leveraging this additional information, our model can outperform prior methods in complex scenarios where text and visual elements are closely intertwined. For instance, consider the scenario illustrated in Fig. 3. When the term "apples" in "piles of apples behind the grapes" is masked, a model lacking precise word-object correlation might falter in predicting the appropriate term. Several other words might yield a semantically consistent sentence, but they would not be accurate in the given visual context.

3.3. Cross-modal Alignment Module

To further improve the performance of our model, we also make a meaningful improvement to the popular cross-modal alignment mechanism proposed by prior work [64]. As depicted in Fig. 4, our cross-modal alignment module (CAM) improves language-image alignment by injecting global contextual prior into image features before performing language-image fusion. CAM first uses pooling oper-

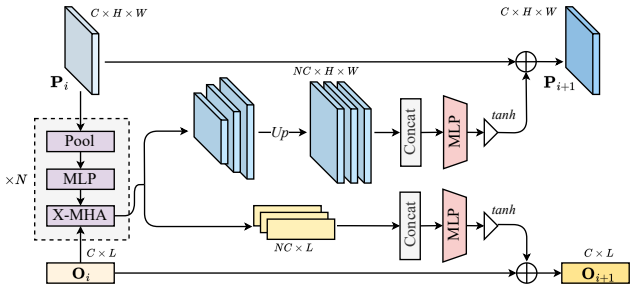


Figure 4. Cross-modal Alignment Module. This module enables bidirectional language-image interaction and addresses granularity mismatches between language and image features, thereby enhancing segmentation accuracy for RIS. X-MHA denotes bidirectional cross-modal multi-head attention. $\mathbf{P}_i$ and $\mathbf{P}_{i+1}$ denote input and output image features, whereas $\mathbf{O}_i$ and $\mathbf{O}_{i+1}$ denote input and output language features. Up denotes upsampling.

ations with different window sizes to generate K feature maps of different pyramid scales. Then, each of these feature maps passes through a 3-layer MLP to better extract global information, before cross-attending with the opposite modality. After that, all the output features are upsampled to the original feature map via bilinear interpolation and concatenated along the channel dimension. A 2-layer MLP is subsequently used to to reduce the channel dimension of this concatenated feature back to the original one. To prevent the multi-modal signal from overwhelming the original signal, a gate with Tanh nonlinearity is used to modulate the final output. Finally, this post-gate feature is added back to the input feature before being passed to the next stage of the image or language encoder. We split language encoder into 4 stages with an equal number of layers and add CAM to the end of every stage of image and language encoder.

Mathematically, CAM can be represented as follows: Let $\mathbf{T}_i$ and $\mathbf{I}_i$ be the text/image input to each stage of the language and image encoder. At each stage,

$$\mathbf{O}_i = \text{LanguageStage}(\mathbf{T}_i),\ \mathbf{P}_i = \text{ImageStage}(\mathbf{I}_i), \tag{5}$$

$$\mathbf{P}_i^k = \text{MLP}_k(\text{Pool}_k(\mathbf{P}_i)), \tag{6}$$

$$\mathbf{O}_{i,p2t}^K, \mathbf{P}_{i,t2p}^k = \text{X-MHA}_k(\mathbf{O}_i, \mathbf{P}_n^k), \tag{7}$$

$$\mathbf{O}_{i,p2t} = \text{Concat}([\mathbf{O}_{i,p2t}^i, ..., \mathbf{O}_{i,p2t}^N], \tag{8}$$

$$\mathbf{P}_{i,t2p} = \text{Concat}([\text{Up}(\mathbf{P}_{i,t2p}^1, ..., \text{Up}(\mathbf{P}_{i,t2p}^N)], \tag{9}$$

$$\mathbf{O}_{i+1} = \mathbf{O}_i + \tanh(\text{MLP}(\mathbf{O}_{i,p2t})), \tag{10}$$

$$\mathbf{P}_{i+1} = \mathbf{P}_i + \tanh(\text{MLP}(\mathbf{P}_{i,t2p})), \tag{11}$$

where Up denotes upsampling and X-MHA [40] denotes bidirectional cross-modal multi-head attention.

CAM enhances cross-modal alignment by enabling bidirectional language-image interaction, which stands in contrast to the widely-used one-way language to image alignment module proposed by LAVT [64]. Moreover, CAM adopts a pyramid pooling technique to utilize multi-scale average-pooled image features. This technique adeptly resolves the granularity mismatch issue by capturing image features at multiple scales, allowing our network to handle the varied levels of detail present in language descriptions. This is particularly beneficial for RIS, where the model must accurately interpret and segment according to a diverse range of descriptive queries.

3.4. Cross-modal Alignment Loss

On top of that, similar to previous works [61, 71], we also use cross-modal alignment loss to explicitly align language and image features. Our cross-modal alignment loss (CAL) is holistic and consider both pixel-to-pixel ($\mathcal{L}_{\text{P2P}}$) and pixel-to-text ($\mathcal{L}_{\text{P2T}}$) consistency.

Mathematically, CAL is computed as follows: Given language feature $\mathbf{T} \in \mathbb{R}^{M \times D}$ produced by the language encoder and final pixel decoder mask feature $\mathbf{I} \in \mathbb{R}^{C_L \times H_L \times W_L}$ with $|\mathcal{P}|$ positive pixel features, $|\mathcal{N}|$ negative pixel features, let $\mathbf{I}_i^+$ be the i^{th} pixel feature in the positive set $\mathcal{P}$, $\mathbf{I}_j^-$ be the j^{th} pixel feature in the background set $\mathcal{N}$ and $\mathbf{T}_k$ be the k^{th} language token, then

$$\mathcal{L}_{\text{CAL}} = \mathcal{L}_{\text{P2P}} + \mathcal{L}_{\text{P2T}} \tag{12}$$

$$\begin{aligned}\mathcal{L}_{\text{P2P}} = &-\frac{1}{|\mathcal{P}|}\sum_i^{|\mathcal{P}|}\frac{e^{\mathbf{I}_i^+ \cdot \mathbf{I}_{\text{avg}}^+/\tau_1}}{e^{\mathbf{I}_i^+ \cdot \mathbf{I}_{\text{avg}}^+/\tau_1} + \sum_j^{|\mathcal{N}|} e^{\mathbf{I}_i^+ \cdot \mathbf{I}_j^-/\tau_1}} \\ &+ -\frac{1}{|\mathcal{N}|}\sum_j^{|\mathcal{N}|}\frac{e^{\mathbf{I}_j^- \cdot \mathbf{I}_{\text{avg}}^-/\tau_1}}{e^{\mathbf{I}_j^- \cdot \mathbf{I}_{\text{avg}}^-/\tau_1} + \sum_i^{|\mathcal{P}|} e^{\mathbf{I}_j^- \cdot \mathbf{I}_i^+/\tau_1}},\end{aligned} \tag{13}$$

$$\mathcal{L}_{\text{P2T}} = -\frac{1}{|\mathcal{P}|}\sum_i^{|\mathcal{P}|}\frac{e^{\mathbf{I}_i^+ \cdot \mathbf{T}_{\text{avg}}/\tau_2}}{e^{\mathbf{I}_i^+ \cdot \mathbf{T}_{\text{avg}}/\tau_2} + \sum_j^{|\mathcal{N}|} e^{\mathbf{I}_i^+ \cdot \mathbf{I}_j^-/\tau_2}}, \tag{14}$$

where $\mathbf{I}_{\text{avg}}^+ = \frac{1}{|\mathcal{P}|}\sum_i^{|\mathcal{P}|}\mathbf{I}_i^+$ and $\mathbf{I}_{\text{avg}}^- = \frac{1}{|\mathcal{N}|}\sum_j^{|\mathcal{N}|}\mathbf{I}_j^-$ are the average pooled positive and negative pixel features, $\mathbf{T}_{\text{avg}} = \text{proj}(\frac{1}{M}\sum_m^M \mathbf{T}_k)$ is the average pooled and linearly projected word feature and τ_1, τ_2 are hyper-parameters that affect the sharpness of the probability distribution. Note that all language and image features are L2-normalized before any dot product computation, but not explicitly shown in the equations above for brevity.

CAL differs from alignment losses used in prior works [61, 71] by holistically integrating both pixel-to-pixel and pixel-to-text alignments within a single cohesive system. Precise pixel-to-pixel alignment ensures that segmentation outputs have accurate shapes and boundaries, whereas precise pixel-to-text alignment enables our model to correctly associate textual descriptions with their matching image regions. This dual alignment mechanism allows our model to effectively parse and interpret the nuanced interplay between image details and language cues, leading to more accurate and contextually relevant segmentation outputs.

	Method	Backbone	RefCOCO (easy)			RefCOCO+ (medium)			G-Ref (hard)		
			val	test A	test B	val	test A	test B	val (U)	test (U)	val (G)
Single Dataset	VLT [14]	Darknet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76
	ReSTR [34]	ViT-B-16	67.22	69.30	64.45	55.78	60.44	48.27	-	-	54.48
	CRIS [61]	ResNet-101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
	LAVT [64]	Swin-B	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
	VPD [72]	Swin-B	73.46	-	-	63.93	-	-	63.12	-	-
	CoupAlign [71]	Swin-B	74.70	77.76	70.58	62.92	68.34	56.69	62.84	62.22	-
	PVD [10]	Swin-B	74.82	77.11	69.52	63.38	68.60	56.92	63.13	63.62	61.33
	SADLR [65]	Swin-B	74.24	76.25	70.06	64.28	69.09	55.19	63.60	63.56	61.16
	MCRES [62]	Swin-B	74.92	76.98	70.84	64.32	69.68	56.64	63.51	64.90	61.63
	ReLA [45]	Swin-B	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97	62.70
	MagNet (Ours)	Swin-B	**75.24**	**78.24**	**71.05**	**66.16**	**71.32**	**58.14**	**65.36**	**66.03**	**63.13**
Multiple / Extra Datasets	SEEM† [75]	Focal-T	-	-	-	-	-	-	65.7	-	-
	LISA-7B† [38]	SAM-H	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	-
	PolyFormer† [46]	Swin-B	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05	-
	MagNet‡ (Ours)	Swin-B	**76.55**	**78.27**	**72.15**	**68.10**	**73.64**	**61.81**	**67.79**	**69.29**	-

Table 1. Comparison with SOTA methods using the oIoU metric. *Single dataset* refers to strictly following the predefined train/test splits of the original RefCOCO, RefCOCO+ and G-Ref datasets. *Multiple datasets* refers to combining the train splits from these 3 datasets with test images removed to prevent data leakage. *Extra datasets* refers to using additional data beyond RefCOCO, RefCOCO+ and G-Ref. † indicates models that use extra datasets. ‡ indicates that our model only uses multiple datasets. **Bold** indicates best.

3.5. Loss Function

Our loss function is a weighted combination of the following 4 different losses:

$$\mathcal{L} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{CAL}}\mathcal{L}_{\text{CAL}} + \lambda_{\text{grounding}}\mathcal{L}_{\text{grounding}}, \tag{15}$$

with $\lambda_{\text{BCE}} = 2.0$, $\lambda_{\text{Dice}} = 2.0$, $\lambda_{\text{CAL}} = 0.5$, and $\lambda_{\text{grounding}} = 1.0$ for all our experiments.

4. Experiments

In this section, we first describe the datasets and evaluation metrics (Sec. 4.1). Then, we compare our method with SOTA RIS methods (Sec. 4.2). Finally, we show some visualization results (Sec. 4.3) and ablate our proposed method (Sec. 4.4). Due to space limitation, exact implementation details of our method are relegated to the Supplementary.

4.1. Datasets and Evaluation Metrics

We evaluate our proposed method on three standard benchmark datasets: RefCOCO [67], RefCOCO+ [67], and G-Ref [49, 50] using three commonly used metrics: overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision values at 0.5, 0.7, and 0.9 IoU threshold levels (P@X). More details regarding these datasets and metrics can be found in the Supplementary.

4.2. Main Results

In Tab. 1, we evaluate MagNet against other SOTA methods on RefCOCO [67], RefCOCO+,[67] and G-Ref [49, 50] datasets using the oIoU metric. Under the *single dataset* setting, MagNet is the first method that consistently outperforms all previous methods on all evaluation subsets of these datasets. Previous methods usually overfit to one of these benchmarks and perform worse in others. Remarkably, on RefCOCO, MagNet outperforms the very recent SOTA RIS method ReLA [45] by considerable margins of **1.42**, **1.76**, and **0.87** points on the validation, testA, and testB subsets, respectively. To have a more comprehensive evaluation of our method, we also assess MagNet using other metrics and display the results on Tab. 2. As shown, MagNet has much better mIoU and P@X performance than all previous SOTA methods. In particular, our method surpasses previous SOTA methods by **0.32** points on the oIoU metric, **0.91** points on the mIoU metric and **0.89**, **1.25**, **1.41** points on the precision metric at 0.5, 0.7 and 0.9 IoU threshold levels. Under the *multiple / extra datasets* setting, our method also surpasses recent SOTA methods [38, 46] that use large language models [57] or has much slower inference speed, by large margins of up to **2.48** points.

Method	RefCOCO val				
	oIoU	mIoU	P@0.5	P@0.7	P@0.9
LAVT [64]	72.73	74.46	84.46	75.28	34.30
ReLA [45]	73.82	75.61	85.82	77.71	34.99
CoupAlign [71]	74.70	75.49	86.40	77.59	32.40
MCRES [62]	<u>74.92</u>	-	86.23	77.25	35.61
SADLR [65]	74.24	<u>76.52</u>	<u>86.90</u>	<u>78.76</u>	<u>37.36</u>
MagNet (ours)	**75.24**	**77.43**	**87.79**	**80.01**	**38.77**

Table 2. Comparison with SOTA methods on RefCOCO val using oIoU, mIoU and P@X (Precision at IoU threshold value X). **Bold** indicates best and <u>underline</u> indicates second best.

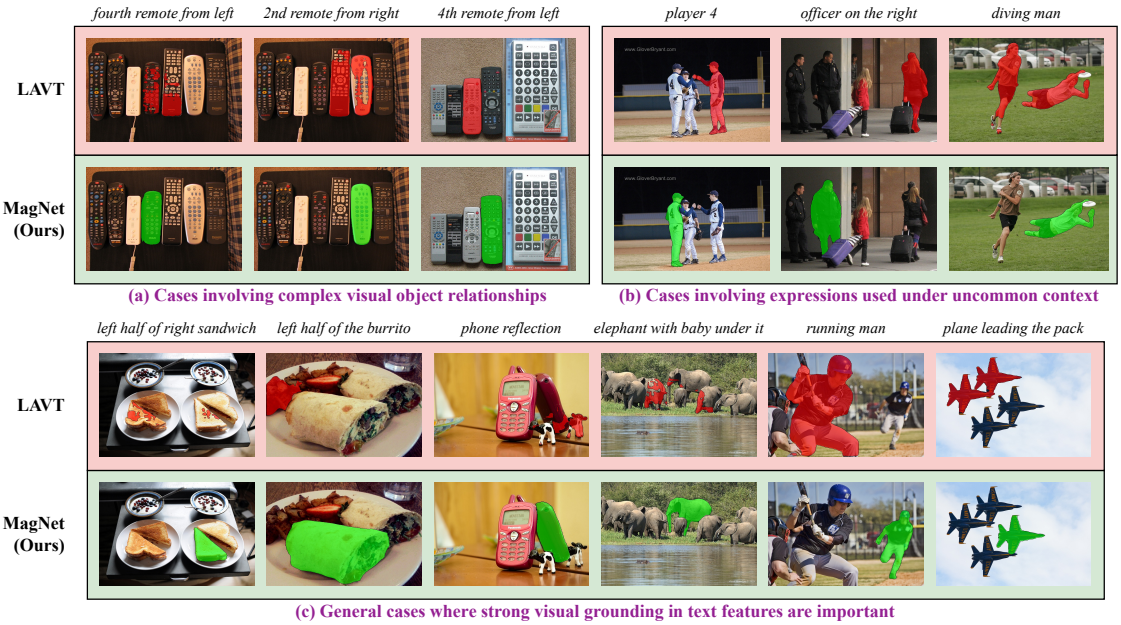


Figure 5. Visualization of MagNet's predictions. Compared to one of the state-of-the-art method, LAVT, our method performs much better in various complex scenerios, suggesting its impressive capability to reason about various complex visual-object relationships.

4.3. Visualizations

In Fig. 5, we show some representative mask predictions of MagNet and LAVT [64] on RefCOCO validation set. Here, we only compare with LAVT because it is a SOTA method that provides reproducible codes and pre-trained weights. MagNet outperforms LAVT in scenes that involve complex visual-object relationships, contain uncommon expressions or require strong visual grounding information. Impressively, MagNet demonstrates ability to grasp complex visual cues such as *reflection*, *leading* and *running*.

4.4. Ablation Studies

In this section, we investigate the effectiveness of all core components of our model. For experimental efficiency, we use a shorter training schedule of 10 epochs and smaller input images of 224×224, causing the results to be different from Tab. 1. Other experimental settings are kept the same. We reproduce the numbers for LAVT [64], ReLA [45] and CRIS [61] using their official codes. All ablations are performed on validation splits of RefCOCO and RefCOCO+.

4.4.1 Ablating Different Aspects of Mask Grounding

Effect on RIS Performance. In Tab. 3(a), we show that both masked language modeling (MLM) and masked vision-language modeling (MaskedVLM) fail to deliver meaningful performance gains. In contrast, our proposed Mask Grounding improves over MLM by encouraging our model to learn fine-grained visual grounding in language features through usage of additional visual and segmentation information. When added to LAVT, Mask Grounding yields a significant performance gains of 1.44 points on RefCOCO and 1.28 points on RefCOCO+.

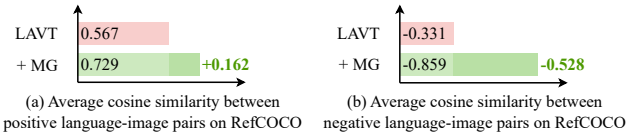


Figure 6. Mask Grounding Improves Language-Image Alignment.

Effect on Language-Image Alignment. Next, we check if Mask Grounding can help to improve language-image alignment in RIS models. Effective alignment is indicated by high feature similarity for matching language-image pairs and low similarity for non-matching pairs. To verify this property, we compare the average normalized cosine similarity for all language-image pairs in the RefCOCO validation dataset before and after Mask Grounding is added to LAVT. Since language and image features have different dimensions, we first train a linear layer with contrastive loss to project language features to the same dimension as image features before computing this metric. This method is similar to linear-probing widely used in self-supervised learning [7, 25]. All other weights are frozen in the process. As illustrated in Fig. 6, Mask Grounding can indeed significantly improve language-image alignment in existing RIS models.

Mask Encoder Design. In Tab. 3(b), we compare passing two different types of mask input to the mask encoder in Mask Grounding. As shown, using center coordinates of masked region gives slightly better performance.

Model	RefCOCO	RefCOCO+
Baseline	67.08	55.98
Baseline + MLM	67.31 (+0.23)	55.69 (-0.29)
Baseline + MaskedVLM	67.33 (+0.25)	56.13 (+0.15)
Baseline + MG	**68.52** (+1.44)	**57.26** (+1.28)

(a) Effectiveness of Mask Grounding over masked language modeling (MLM) and masked vision-language modeling (MaskedVLM).

Model	RefCOCO	RefCOCO+
LAVT [64]	67.08	55.98
LAVT + MG	**68.52** (+1.44)	**57.26** (+1.28)
ReLA [45]	66.23	53.69
ReLA + MG	**67.33** (+1.10)	**55.21** (+1.52)
CRIS [61]	64.67	54.84
CRIS + MG	**65.56** (+0.89)	**56.23** (+1.39)

(d) Universality of Mask Grounding.

Type	RefCOCO	RefCOCO+
None	67.08	55.98
Average	68.19 (+1.11)	56.98 (+1.00)
Center	**68.52** (+1.44)	**57.26** (+1.28)

(b) Comparing different mask encoder input in Mask Grounding. *Center* denotes center coordinates of masked region. *Average* denotes average visual features within masked region.

Pyramid scales	RefCOCO	RefCOCO+
None	67.08	55.98
{1}	67.18 (+0.10)	56.20 (+0.22)
{1,2}	67.44 (+0.36)	56.46 (+0.48)
{1,2,3}	67.79 (+0.71)	56.74 (+0.76)
{1,2,3,6}	**68.06** (+0.98)	**57.04** (+1.06)

(e) Effectiveness of language-image Cross-modal Alignment Module at different scales.

MLP layers	RefCOCO	RefCOCO+
None	67.08	55.98
4	67.14 (+0.06)	57.19 (+1.21)
8	**68.52** (+1.44)	**57.26** (+1.28)
12	66.92 (-0.16)	55.95 (-0.03)

(c) Sensitivity of Mask Grounding's masked token predictor to different MLP layers.

$\mathcal{L}_{\text{P2P}}$	$\mathcal{L}_{\text{P2T}}$	RefCOCO	RefCOCO+
✗	✗	67.08	55.98
✓	✗	67.63 (+0.55)	56.87 (+0.89)
✗	✓	68.27 (+1.19)	57.22 (+1.24)
✓	✓	**68.44** (+1.36)	**57.61** (+1.63)

(f) Effectiveness of language-image Cross-modal Alignment Loss.

Table 3. Ablation Experiments. All experiments are run with a shorter training schedule of 10 epochs, causing the results here to be different from the main results. Rows marked in gray indicate options that are used in the main results. *MG* denotes Mask Grounding.

Mask Token Predictor Design. In Tab. 3(c), we evaluate the sensitivity of Mask Grounding's masked token predictor to different MLP layers. We observe that a sufficiently deep masked token predictor is important for good performance. As shown, performance is the best when 8 layers are used. When more layers are added, performance slightly drops, as the model starts to overfit to the auxiliary task.

Universality of Mask Grounding. In Tab. 3(d), we show that Mask Grounding is also compatible with other representative RIS methods. As shown, when Mask Grounding is added, on RefCOCO and RefCOCO+, we can obtain a performance gain of 0.89 and 1.29 points for CRIS and a performance gain of of 1.10 and 1.52 points for ReLA.

4.4.2 Ablating Other Components of Our Method

Effectiveness of CAM. Cross-modal Alignment Module (CAM) improves language-image alignment by injecting global contextual prior into image features. As shown in Tab. 3(e), when CAM is used, we can improve RefCOCO's and RefCOCO+'s oIoU by 0.98 and 1.06 points respectively. Additionally, Tab. 3(e) also shows that using more pyramid scales is helpful in boosting performance.

Effectiveness of CAL. Cross-modal Alignment Loss (CAL) provides additional pixel-to-pixel ($\mathcal{L}_{\text{P2P}}$) and pixel-to-text ($\mathcal{L}_{\text{P2T}}$) alignment supervision to further reduce language-image modality gap. As shown in Tab. 3(f), both $\mathcal{L}_{\text{P2P}}$ and $\mathcal{L}_{\text{P2T}}$ alone can bring noticeable oIoU improvement on RefCOCO and RefCOCO+. When the both are added together, we can surpass the baseline by 1.36 points on RefCOCO and 1.63 points on RefCOCO+.

Compatibility of all MagNet components. As shown in Fig. 7, all components of MagNet are highly compatible as they progressively improves the LAVT's performance when added incrementally. When all components are added, we can improve LAVT by 3.15 points on RefCOCO+.

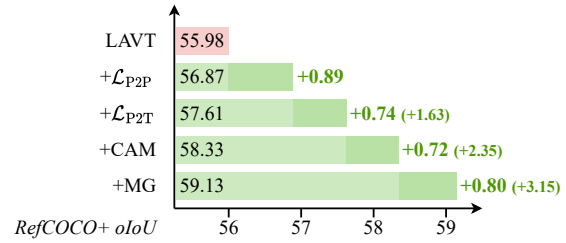


Figure 7. Compatibility of MagNet components.

5. Conclusion

In this paper, we present Mask Grounding, an novel method designed to enhance RIS by teaching our model to predict randomly masked textual tokens based on their surrounding textual, visual and segmentation information. This task requires our model to learn fine-grained visual-textual object correspondence, thus learning visual grounding in the process. When plugged into existing RIS algorithms, Mask Grounding can improve their performance consistently. To holistically address the modality gap, we also design a cross-modal alignment loss and an accompanying alignment module. When all these techniques are used together, our newly proposed MagNet achieves SOTA performance in all RIS benchmarks. We believe that Mask Grounding can also be used in other multi-modal dense prediction tasks and will explore that in future work.

Acknowledgements. We thank Wan Ding and Kai Ding for their kind support in this project. This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0140407, the National Natural Science Foundation of China under Grants 62321005 and 62276150, and the THU-Bosch JCML.

References

[1] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh. Interactive text2pickup networks for natural language-based human–robot collaboration. *IEEE Robotics and Automation Letters*, 2018. 1

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 1

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 3, 4

[4] D. Chen, S. Jia, Y. Lo, H. Chen, and T. Liu. See-through-text grouping for referring image segmentation. In *ICCV*, 2019. 3

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 1

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[7] X. Chen and K. He. Exploring simple siamese representation learning. In *CVPR*, 2021. 7

[8] B. Cheng, M. Collins, Y. Zhu, T. Liu, T. Huang, A. Hartwig, and L. Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 1

[9] M. Cheng, S. Zheng, W. Lin, V. Vineet, P. Sturgess, N. Crook, N. J. Mitra, and P. H. Torr. Imagespirit: Verbal guided image parsing. *ACM Transactions on Graphics (ToG)*, 2014. 1

[10] Z. Cheng, K. Li, P. Jin, X. Ji, L. Yuan, C. Liu, and J. Chen. Parallel vertex diffusion for unified visual grounding. In *AAAI*, 2024. 3, 6

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv:2210.11416*, 2022. 3

[12] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 1

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NACL: HLT*, 2019. 3, 4

[14] H. Ding, C. Liu, S. Wang, and X. Jiang. Vision-language transformer and query generation for referring segmentation. In *CVPR*, 2021. 1, 2, 3, 6

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[16] G. Feng, Z. Hu, L. Zhang, and H. Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021. 3

[17] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 3

[18] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[19] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang. Dynamic neural networks: A survey. *IEEE TPAMI*, 2021. 3

[20] Y. Han, Gao. Huang, S. Song, L. Yang, Y. Zhang, and H. Jiang. Spatially adaptive feature refinement for efficient inference. *IEEE TIP*, 2021.

[21] Y. Han, Z. Yuan, Y. Pu, C. Xue, S. Song, G. Sun, and G. Huang. Latency-aware spatial-wise dynamic networks. *NeurIPS*, 2022. 3

[22] Y. Han, D. Han, Z. Liu, Y. Wang, X. Pan, Y. Pu, C. Deng, J. Feng, S. Song, and G. Huang. Dynamic perceiver for efficient visual recognition. In *ICCV*, 2023. 3

[23] Y. Han, Z. Liu, Z. Yuan, Y. Pu, C. Wang, S. Song, and G. Huang. Latency-aware unified dynamic networks for efficient image recognition. *arXiv:2308.15949*, 2023. 3

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[25] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 7

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3

[27] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[28] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1, 3

[29] Y. Hu, Qi Wang, W. Shao, E. Xie, Z. Li, J. Han, and P. Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*, 2023. 1, 3

[30] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu. Bidirectional relationship inferring network for referring image localization and segmentation. In *CVPR*, 2020. 3

[31] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020. 3

[32] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020. 3

[33] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3, 4

[34] N. Kim, D. Kim, C. Lan, W. Zeng, and S. Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, 2022. 6

[35] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 3

[36] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *CVPR*, 2019. 1

[37] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto. Masked vision and language modeling for multi-modal representation learning. *ICLR*, 2023. 3, 4

[38] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv:2308.00692*, 2023. 6

[39] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3

[40] L. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J. Hwang, K. Chang, and J. Gao. Grounded language-image pre-training. In *CVPR*, 2022. 5

[41] R. Li, K. Li, Y. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 3

[42] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[43] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler. Editgan: High-precision semantic image editing. In *NeurIPS*, 2021. 1

[44] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017. 3

[45] C. Liu, H. Ding, and X. Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023. 2, 3, 6, 7, 8

[46] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *CVPR*, 2023. 3, 6

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[48] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[49] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2, 6

[50] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2, 6

[51] Z. Ni, Y. Wang, R. Zhou, J. Guo, J. Hu, Z. Liu, S. Song, Y. Yao, and G. Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, 2024. 3

[52] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *CVPR*, 2021. 1

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[54] H. Shi, H. Li, F. Meng, and Q. Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018. 3

[55] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[56] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 3

[57] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 3, 4, 6

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[59] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. V. Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 3

[60] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. In *CVPR*, 2023. 3

[61] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 2, 5, 6, 7, 8

[62] L. Xu, M. H. Huang, X. Shang, Z. Yuan, Y. Sun, and J. Liu. Meta compositional referring expression segmentation. In *CVPR*, 2023. 3, 6

[63] Y. Yan, X. He, W. Wan, and J. Liu. Mmnet: Multi-mask network for referring image segmentation. *ACM MM*, 2023. 2

[64] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[65] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr. Semantics-aware dynamic localization and refinement for referring image segmentation. In *AAAI*, 2023. 3, 6

[66] L. Ye, M. Rochan, Z. Liu, and Y. Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019. 3

[67] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 6

[68] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 3

[69] A. Zareian, K. D. Rosa, D. Hu, and S. Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 2

[70] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv:2205.01068*, 2022. 3

[71] Z. Zhang, Y. Zhu, J. Liu, X. Liang, and W. Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. In *NeurIPS*, 2022. 1, 2, 3, 5, 6

[72] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 2, 3, 6

[73] X. Zhao, R. Vemulapalli, P. A. Mansfield, B.Gong, B. Green, L. Shapira, and Y. Wu. Contrastive learning for label efficient semantic segmentation. In *ICCV*, 2021. 3

[74] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, 2022. 3

[75] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. Lee. Segment everything everywhere all at once. *NeurIPS*, 2023. 6