LoA-Trans: Enhancing Visual Grounding by Location-Aware Transformers

Ziling Huang[1,2] and Shin'ichi Satoh[2,1]

[1] The University of Tokyo, Tokyo, Japan
[2] National Institute of Informatics, Tokyo, Japan
{huangziling,satoh}@nii.ac.jp

Abstract. Given an image and text description, visual grounding will find target region in the image explained by the text. It has two task settings: referring expression comprehension (REC) to estimate bounding-box and referring expression segmentation (RES) to predict segmentation mask. Currently the most promising visual grounding approaches are to learn REC and RES jointly by giving rich ground truth of both bounding-box and segmentation mask of the target object. However, we argue that a very simple but strong constraint has been overlooked by the existing approaches: given an image and a text description, REC and RES refer to the same object. We propose **Lo**cation **A**ware **Trans**former (LoA-Trans) making this constraint explicit by a *center prompt*, where the system first predicts the center of the target object by Location-Aware Network, and feeds it as a common prompt to both REC and RES. In this way, the system constrains that REC and RES refer to the same object. To mitigate possible inaccuracies in center estimation, we introduce a query selection mechanism. Instead of random initialization queries for bounding-box and segmentation mask decoding, the query selection mechanism generates possible object locations other than the estimated center and use them as location-aware queries as a remedy for possible inaccurate center estimation. We also introduce a TaskSyn Network in the decoder to better coordination between REC and RES. Our method achieved state-of-the-art performance on three commonly used datasets: Refcoco, Refcoco+, and Refcocog. Extensive ablation studies demonstrated the validity of each of the proposed components.

Keywords: Multi-task Learning · Center Prompt · TaskSyn Network

1 Introduction

Visual Grounding includes two tasks: Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES). REC [1, 6, 7, 13, 18, 20, 30, 31, 35, 38, 39, 42, 44, 45] tends to identify the specific entity or object being referred to in language and estimate its bounding-box within an image while RES [4, 8, 24, 30, 34, 36, 37] involves segmenting the specific region referred to in language within an image. Both tasks are key focus in vision language research, demanding comprehensive understanding within a single modality like image or language and achieving precisely cross-modality alignment between language and image.

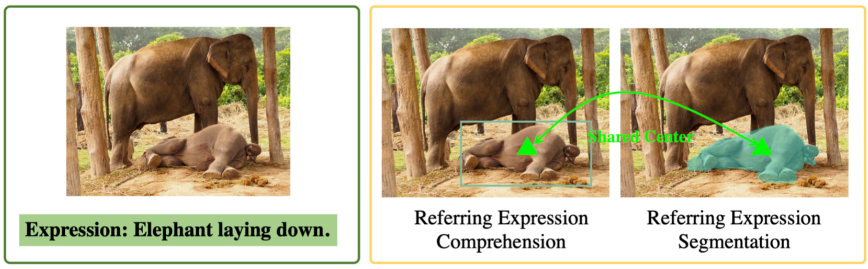

Fig. 1: This figure illustrates the relations between Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES). These two tasks have the same input images and natural language expression. Although their objectives are different, the objects its natural language expression referred to are the same.

Previous works treated REC and RES as separate tasks, designing methods specifically for each. REC can be categorized into two-stage pipeline and one-stage pipeline. Two-stage pipeline [6,7,18,20,35,42,44] generates multiple proposals in the first stage, then finds the most relevant proposal. One stage pipeline [1, 13, 30, 31, 38, 39, 45] regresses the target bounding-box directly. On the other hand, RES mainly focuses on one-stage pipeline [8, 9, 17, 27, 41]. However, there is a growing interest in exploring the synergy between these two tasks through joint learning frameworks for two reasons: (1) By sharing parameters and learning jointly, the model can avoid redundant computations and save training time, leading to more efficient use of computational resources. (2) Integrating REC and RES into a unified model, allows the model to learn richer representations through different objectives. Several methods are proposed with their solutions in joint learning for RES and REC. RefTR [12] utilizes a transformer for visual-language alignment and proposes separated heads for bounding-box prediction and segmentation map estimation. It allows the model to learn richer representations by different optimization objectives of REC and RES. To better integrate the two tasks, in Polyformer [19] and SeqTR [46], the segmentation mask is represented as a sequence of discrete coordinate tokens. In this way, REC and RES are unified as point prediction problems with the same optimization objective.

Although these methods achieve good performance through joint training, the two tasks communicate in the decoder by self-attention mechanism, which suffers some drawbacks: (1) The self-attention mechanism is responsible for information exchange only but does not explicitly assume that REC and RES address the same object. Although the objectives of REC and RES are different, the objects its natural language expression referred to are the same when they have the same input images and language expression, as shown in Fig. 1. (2) In self-attention mechanisms, information sharing happens between all queries, without any specific coordination between tasks. The information exchange between two tasks is very important. For example, the segmentation map can provide size information which obviously is informative for bounding-box estimation as well. Considering that, we present **Lo**cation **A**ware **Trans**former (LoA-Trans) by introducing a *center prompt*: a center point of the target object to be fed to both REC and RES as a common location indicator, to explicitly constrain the targeted object for both REC and RES is the same. TaskSyn Network further attains better task

information exchange. The center prompt acts like a spotlight, directing the decoder's focus to the object mentioned in the language, resulting in better bounding box and segmentation mask estimation. However, to mitigate potential inaccuracies in center prompt estimation, we introduce a query selection mechanism as a backup strategy. Instead of using random initialization queries for bounding-box and segmentation mask decoding, we develop a query selection mechanism to pick location-aware queries for bounding-box estimation and segmentation. By guiding the decoder's focus towards the object described in the language, our model can estimate the bounding box and segmentation mask of referred objects successfully. Furthermore, the TaskSyn Network is designed to improve communication between the two tasks. Because REC and RES have different goals but their information still benefits each other, the the design of TaskSyn Network helps them process information separately while still allowing them to share useful insights.

In summary, our contributions in LoA-Trans are threefolds:

- We introduce a center prompt to explicitly constrain that both RES and REC refer to the same object. We also experimentally showed that the center estimation is relatively insensitive to the quality of the final outputs.
- Instead of using random initialization queries for bounding-box and segmentation mask decoding, we develop a query selection mechanism to pick initial location-aware queries for better bounding-box and segmentation estimation results.
- Recognizing the distinct and complementary nature of REC and RES, we propose a TaskSyn Network for effective information exchange between two tasks.

Our methods achieved state-of-the-art performance on three commonly used datasets: Refcoco [43], Refcoco+ [43] hand Refcocog [26]. Extensive ablation studies demonstrated the validity of each of the proposed components.

2 Related Work

Referring Expression Comprehension (REC). The early methods in REC mainly focused on two-stage pipeline [6, 7, 18, 20, 35, 42, 44]. The two-stage pipeline employs the most popular detection framework e.g., Faster RCNN [5] to obtain object proposals. Then, all object proposals will be ranked by designed similarity score calculation methods. To avoid detection error, one-stage pipeline [1, 13, 30, 31, 38, 39, 45] attracts more attention recently. The transformer-like structure [1, 10] is proven to be effective in this task.

Referring Expression Segmentation (RES). Compared with REC, most RES methods employed a one-stage pipeline. Early RES focuses on better vision and language alignment by designing different types of attention [8, 9, 17, 27, 41]. Not so much work implements methods like those above anymore because a one-stage interaction between visual and language information is not enough to predict accurate pixel-wise masks. The more recent second type of method aims at achieving better language and visual alignment by breaking down complex problems into cascade multiple steps. CGAN [24] and CRIS [34] have a cascade framework in which language features can fuse many times with multi-modality features output from the previous layer. EFN [4] and LAVT [37]

show that better cross-modal alignments can be achieved through the early fusion of linguistic and visual features. VLT [3] generates multiple sets of word attention weights to represent different understandings.

Multi-Task Referring Grounding. The target of Multi-Task Referring Grounding is to address RES and REC tasks jointly. MCN [25] used consistent energy maximization to bridge two tasks together. RefTR [12] utilizes a transformer for visual-language alignment and proposes separated decoder heads for bounding-box prediction and segmentation map estimation. It allows the model to learn richer representations by different optimization objectives. To better integrate the two tasks, In Polyformer [19] and SeqTR [46], the segmentation mask is represented as a sequence of discrete coordinate tokens. In this way, REC and RES are unified as point prediction problems with the same optimization objective. Different from these methods, our proposed LoA-Trans explicitly points out that REC and RES refer to the same object when provided with identical inputs and perform specific information exchange between tasks.

3 Our Method: LoA-Trans

In this section, we will introduce our proposed LoA-Trans, including visual encoder, text encoder, decoder, and training objectives. Our LoA-Trans is based on Deformable DETR [47]. Fig. 2 shows the overall architecture of the method. We first perform visual-language alignment by early fusion. The aligned multi-modality feature is used to estimate the initial object location by Location-Aware Network as the estimated center prompt, to be fed to Queries Decoder as a common prompt for both REC and RES. As the remedy for possible errors caused by Location-Aware Network, Mask Queries Selection will generate other possible target object candidates as Bounding-Box and Mask Queries.

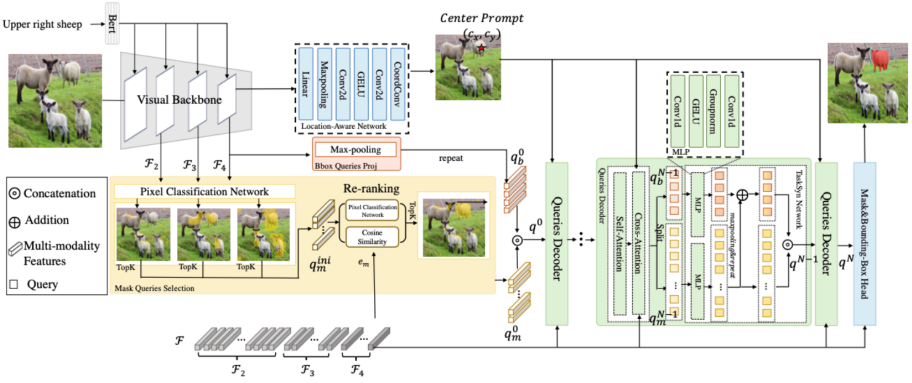

Fig. 2: Overall framework of proposed methods. The early fusion is for visual-language alignment. The last layer feature maps are used for the center prompt and pixel classification network. After re-ranking, the multi-modality features most related to the referred object are selected, marked with yellow. The proposed TaskSyn Network passes information of segmentation queries to bounding box queries to achieve information exchange. Note, after query selection, the multi-modality features $\mathcal{F}$ are sent into six layers transformer encoder before decoder.

3.1 Image & Text Feature Extraction and Fusion

To estimate the target segmentation mask and bounding box, we need to achieve visual and language alignment first. We assume that an image $\mathcal{I}$ and a free-form language expression with L words are given. Our approach employs the Swin Transformer [21] as the visual encoding backbone to transform an image $\mathcal{I}$, into a set of multi-level features. We represent these features as $\mathcal{V} = \{\mathcal{V}_i\}_{i=1}^{4}$ with distinct dimensions in terms of height H_i, width W_i, and channels C_i. We assume $H_{i+1} = \lfloor W_i/2 \rfloor$ and so on in our implementation. These multi-level features correspond to different stages within the Swin Transformer, offering a comprehensive analysis of the image. Moreover, we adopt BERT [2], a robust language encoder, as our language encoding backbone. The extracted language features are denoted as $\varepsilon = \{e_l\}_{l=1}^{L}$.To distill a global language representation, we average the features of all words, resulting in $e_m \in \mathbb{R}^C$, where C is the channel dimension.

In this work, we use an early fusion approach to produce multi-modality features similar to LAVT [37]. It begins with aligning language features $\varepsilon = \{e_l\}_{l=1}^{L}$ with the first layer of visual features $\mathcal{V}_1$ using a cross-attention mechanism [33]. This creates first-layer attention maps, pinpointing areas in the image linked to text. Attention maps then add with the original image features $\mathcal{V}_1$, forming enhanced multi-modality features $\mathcal{F}_1$. These multi-modality features are then advanced to the next layer to produce $\mathcal{V}_2$. Then, we perform a similar process to create $\mathcal{F} = \{\mathcal{F}_i\}_{i=2}^{4}$. All these operations are performed inside the visual encoding backbone. This step-by-step integration ensures the gradual and thorough merging of both modalities, leading to a deeper understanding of image and language.

3.2 Decoder

REC and RES are two very similar tasks where they share the same inputs and targeted objects, but their output style differs. The LoA-Trans is built on Deformable DETR [47], whose inputs include image features, reference points, and queries. In our proposed LoA-trans, we first estimate a center prompt to indicate the location of the target object, where the center prompt acts as a reference point to direct the decoder focus to the object mentioned by the language. Then, we developed a query selection mechanism to pick location-aware queries instead of random initial queries as in Deformable DETR for bounding-box and segmentation estimation. Finally, we modify the decoder of Deformable DETR by proposing TaskSyn Network. Besides information exchange within queries via self-attention mechanism, our TaskSyn Network enables updated segmentation queries add to updated bounding-box queries to achieve communication between two tasks precisely. The details are illustrated in Fig 2.

Location-Aware Network. In LoA-Trans, since REC and RES address the same object with the same inputs, we introduce a center prompt to highlight the targeted object for both RES and REC are the same. The center prompt acts as a reference point to direct the decoder's focus to the object mentioned in the language. To estimate the center prompt, we design a Location-Aware Network. The Location-Aware Network utilizes multi-modality features in the last layer $\mathcal{F}_4$ and outputs coordinates $\mathbf{c} \in \mathbb{R}^2$ as

shown in Equation 1, indicating where the objects mentioned by the language are likely located.

$$c = Linear(Maxpooling(Conv(GELU(Conv(CoordConv(F_4))))))) \quad (1)$$

Bounding-Box Queries. Instead of random initial queries, we hope bounding-box queries in Decoder are particularly effective at focusing on multi-modality features that are relevant to the object being referred to. To achieve this, we adopt queries that are specifically linked to the target object. After multi-layer early fusion, the target object shows the high-level response in the final multi-modality feature map $\mathcal{F}_4$. Unlike segmentation mask estimation, bounding-box estimation does not require detailed boundary information. Therefore, for simplicity, we employ max-pooling in the spatial dimension of $\mathcal{F}_4$ to extract highly responsive bounding-box queries q_b, as shown in Fig. 2. After the multi-layer LoA-trans decoder, the q_b already has enough information for bounding-box decoding. For bounding-box decoding, we use bounding-box head: multi-layer MLPs, project queries dimension C into coordinate dimension 2 as four direction relative displacement $\Delta*$, we decode the bounding-box as left $b_l \in \mathbb{R}^2$, bottom $b_b \in \mathbb{R}^2$, right $b_r \in \mathbb{R}^2$, up $b_u \in \mathbb{R}^2$, width $b_w \in \mathbb{R}^1$ and height $b_h \in \mathbb{R}^1$ based on center prompt $c \in \mathbb{R}^2$ and estimated relative displacement $\Delta*$. The final estimated bounding box is represented by Equation 2.

$$\begin{cases} b_l = c + \Delta_l \\ b_b = c + \Delta_b \\ b_r = c + \Delta_r \\ b_u = c + \Delta_u \end{cases} \Longrightarrow \begin{cases} c = [\frac{(b_r^x + b_l^x)}{2}, \frac{(b_b^y + b_u^y)}{2}] \\ b_w = b_r^x - b_l^x \\ b_h = b_b^y - b_u^y \end{cases} \quad (2)$$

Mask Queries. Because mask decoding requires boundary information, using max-pooling as bounding box queries alone may not be sufficient for detailed information. Furthermore, recognizing the potential for inaccuracies in center prompt estimation, we tend to fix this problem by selecting location-aware queries to help find the correct target. Considering that, we proposed a mask query selection mechanism. In the mask query selection mechanism, we select queries from multi-level multi-modality features $\mathcal{F}_i$. This selection process is initiated by training a pixel classification network on the last layer of multi-modality features $\mathcal{F}_4$. Following this, the multi-modality features from different layers - $\mathcal{F}_4$, $\mathcal{F}_3$, and $\mathcal{F}_2$ - are passed through pixel classification network. The top-performing features are chosen as queries for each respective layer based on their performance classification score S_c, the selected features are concatenated together and represent as $q_m^{ini} \in \mathbb{R}^{3K \times C}$. This method ensures that the most relevant and informative features across various levels of the network are utilized for precise mask query selection. For further refinement, we implement a re-ranking step. It starts with calculating cosine similarity between q_m^{ini} and global language description features e_m. We combine cosine similarity score with classification scores to perform final re-ranking using Equation 3, and select the top K multi-modality features as the final mask queries $q_m \mathbb{R}^{K \times C}$. After multi-layer refinement in the decoder, the q_m is passed to the segmentation head for final segmentation mask decoding. The details of the segmentation head can be found in the supplementary material.

$$q_m = TopK(S_c(q_m^{ini}) \cdot S_{cos}(q_m^{ini})) \quad (3)$$

where S_{cos} is the score of cosine similarity.

The Decoder is to help the queries to take more related information from multi-modality features for segmentation mask and bounding-box decoding. Previous methods and Deformable DETR [47] perform queries information exchange by self-attention mechanism. In self-attention mechanisms, information sharing happens between all queries, without any specific coordination between tasks and ignoring the fact that REC and RES have different goals. To enable precise task information communication, we proposed TaskSyn Network, as shown in Fig. 2. The proposed TaskSyn Network first utilizes two MLPs to update bounding-box queries and segmentation queries separately. Then, we apply max-pooling to the updated segmentation queries and merge them with the bounding-box queries by addition. This is based on the insight that segmentation mask details can provide crucial size information for bounding box decoding. Consequently, through this decoding process, both the bounding box and segmentation mask queries acquire a wealth of detailed information, equipping them for more effective and accurate object details identification.

Queries in the Decoder aim to form representations of objects by attending to their corresponding spatial locations in the feature maps. To enhance the information captured by queries, we utilize multi-level multi-modality features $\mathcal{F} = \{\mathcal{F}_i\}_{i=2}^{4}$ for Decoder. These features are then flattened and concatenated along the spatial dimension, resulting in $\mathcal{F} \in \mathbb{R}^{C\times(L_1+L_2+L_3)}$, where $(L_1 + L_2 + L_3)$ represents the total length of the concatenated features, as illustrated in Fig. 2.

3.3 Training Loss

During training, there are four losses: Center Estimation Loss $\mathcal{L}_c$, as claimed in Location-Aware Network, loss for pixel classification network $\mathcal{L}_f$. Bounding-Box Estimation Loss $\mathcal{L}_b$ and Segmentation Loss $\mathcal{L}_s$. The total loss is $\mathcal{L}$ as in Equation 4.

$$\mathcal{L} = \lambda_c\mathcal{L}_c + \lambda_f\mathcal{L}_f + \lambda_b\mathcal{L}_b + \lambda_s\mathcal{L}_s \tag{4}$$

In our work, the $\mathcal{L}_c$ loss is L1, and the target center is the bounding box center. $\mathcal{L}_b$ contains L1 loss and GIoU Loss [29]. $\mathcal{L}_f$ and $\mathcal{L}_s$ are composed by focal loss [14] and dice loss [28]. The parameters for λ_c, λ_f, λ_b and λ_s are 2.0, 2.0, 5.0, 2.0, respectively.These specific values are chosen to balance the contribution of each loss type to the total loss, thereby guiding the training process effectively toward accurate and reliable model performance.

4 Experiments

Datasets. To test how well our method works, we ran experiments using three widely-used datasets: Refcoco [43], Refcoco+ [43], Refcocog [26]. Refcoco includes 50,000 different objects across 19,994 images, with a total of 142,209 unique natural language descriptions. Each image contains multiple objects, each described by multiple expressions. Refcoco+ differs from Refcoco in that its referring expressions don't include location words, relying solely on attributes to identify objects. The sentences in Refcoco+ tend to be longer. This dataset comprises 141,564 expressions, 49,856 objects,

and 19,992 images, with descriptions generated through a two-player game [11]. Refcocog is the most challenging of the three. Descriptions are generated via Mechanical Turk, resulting in richer and longer descriptions. This dataset contains 104,560 expressions describing 54,822 objects across 25,711 images, with an average expression length of 8.3 words. For the Refcocog dataset, we utilized the UNC partition. Images for all three datasets were sourced from MS-COCO [15].

Evaluation Metric. To assess our model's performance in referred expressions comprehension (REC), we employ Pr@0.5. This metric deems a prediction accurate if it exhibits an overlap exceeding 0.5 with the actual object bounding box. For Referring Expression Segmentation (RES), we evaluate performance through mIoU, measuring alignment with the ground truth mask. Our analysis also encompasses precision at various IoU thresholds (0.5, 0.7, and 0.9), offering a comprehensive view of the model's effectiveness.

Referring Expression Comprehension (Pr@0.5)								
	Refcoco			Refcoco+			Refcocog	
	val	testA	testB	val	testA	testB	val	test
RvG-Tree [6]	75.06	78.61	69.85	63.51	67.45	56.66	66.95	66.51
CM-A-E [20]	78.35	83.14	71.32	68.09	73.65	58.03	67.99	68.67
FAOA [39]	72.54	74.35	68.50	56.81	60.23	49.60	61.33	60.36
TransVG [1]	80.83	83.38	76.94	68.00	72.46	59.24	68.71	67.98
QRNet [40]	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03
RefTR [12]	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40
SeqTR [46]	81.23	85.00	76.08	68.82	75.37	58.78	71.35	71.58
VG-LAW [30]	86.62	89.32	<u>83.16</u>	76.37	81.04	67.50	76.90	76.96
LoA-Trans-S	<u>87.59</u>	<u>90.17</u>	82.98	<u>78.68</u>	<u>83.93</u>	<u>69.83</u>	<u>79.58</u>	<u>79.29</u>
LoA-Trans-B	**87.75**	**90.60**	**84.81**	**79.56**	**84.95**	**71.75**	**80.80**	**80.18**

Table 1: Comparison with state-of-the-art methods for Referring Expression Comprehension on three widely used datasets. The highest Pr@0.5 is marked with **bold**, while the Th second highest Pr@0.5 is marked with <u>underline</u>.

Implementation Details. For the image encoder, we used Swin Transformer [21] Small for LoA-Trans-S and Base for LoA-Trans-B. Bert [2] was used as the text encoder. All experiments of LoA-Trans-S were conducted on four V100 GPUs, while experiments of LoA-Trans-B were conducted on four A100 GPUs. The decoder in our setup has 6 layers. We resized the input images to 640×640. For the Refcoco and Refcoco+ datasets, we set the sentence length to 20, while for Refcocog, it was set to 30. The model was trained using the AdamW [23] optimizer for 100 epochs with a batch size of 32. The starting learning rate was 0.000020, and it was reduced using CosineAnnealingLR [22] with a T_max of 100. The threshold for the segmentation mask was set at 0.35. We train our model separately for each dataset without combining the training subsets. For Refcoco evaluation, we use the Refcoco training dataset, selecting the best-performing model based on the Refcoco validation dataset with highest mIoU score, as in Refcoco+ and Refcocog. Unless specified otherwise, the number of tokens selected for the 2nd, 3rd, 4th, and final stages is 100, 100, 100, and 100, respectively.

Referring Expression Segmentation (mIoU)								
	Refcoco			Refcoco+			Refcocog	
	val	testA	testB	val	testA	testB	val	test
VLT [3]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
LAVT [37]	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
ReLA [16]	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
RefTR [12]	70.56	73.49	66.57	61.08	64.69	52.73	58.73	58.51
SeqTR [46]	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
VG-LAW [30]	75.62	77.51	72.89	66.63	70.38	58.89	65.63	66.08
LoA-Trans-S	76.03	77.90	72.57	67.85	72.21	60.29	67.44	67.97
LoA-Trans-B	**76.66**	**78.60**	**74.17**	**69.40**	**73.59**	**62.90**	**69.01**	**68.77**

Table 2: Comparison with state-of-the-art methods for Referring Expression Segmentation on three widely used datasets. The highest mIoU is marked with **bold**, while the second highest mIoU is marked with underline.

4.1 Main results

Referring Expression Comprehension. In the landscape of Referring Expression Comprehension, the Table 1 presents an insightful comparative analysis of various models across the Refcoco, Refcoco+, and Refcocog datasets including single-task learning models, RvG-Tree [6], CM-A-E [20] which uses two-stage methods, FAOA [39], TransVG [1], QRNet [40] which uses one-stage methods, and multi-task learning methods RefTr [12], SeqTR [46], and VG-LAW [30]. Notably, LoA-Trans-B demonstrates remarkable performance, surpassing earlier single task methods RvG-Tree [6], CM-A-E [20], and FAOA [39], QRNet [40] by significant margins. For example, LoA-Trans-B achieves a score of 90.06 in Pr@0.5, markedly higher than QRNet [40] which is 85.85 in Pr@0.5 in the Refcoco testA dataset. Even when compared to recent advanced multi-task models like RefTr [12], SeqTR [46], LoA-Trans-B maintains a leading position, especially in more challenging datasets like Refcocog. This indicates not only the effectiveness of the LoA-Trans approach but also its adaptability and robustness in diverse image comprehension contexts, setting new benchmarks in the field.

Referring Expression Segmentation. The provided Table 2, showcasing a range of models for Referring Expression Segmentation across the Refcoco, Refcoco+, and Refcocog datasets, highlights the significant advancements of the LoA-Trans-B model. The models compared include single-task learning methods: VLT [3], LAVT [37], ReLA [16] and multi-task learning methods: RefTR [12], SeqTR [46], VG-LAW [30]. The LoA-Trans-B model not only outperforms earlier approaches like VLT [3] and LAVT [37] but also shows notable improvement over contemporary methods such as ReLA [16], and VG-LAW [30]. For instance, in the challenging Refcoco+ testB, LoA-Trans-B surpasses state-of-the-art method VG-LAW [30] by nearly 5% in mIoU, illustrating its robust segmentation abilities. Additionally, the progression from LoA-Trans-S to LoA-Trans-B within the same series indicates substantial model enhancements, particularly in the more complex Refcocog dataset. This overall performance suggests that LoA-Trans-B is setting new benchmarks in the field, demonstrating a blend of adaptability, accuracy, and advanced skills that are pivotal for future developments in Referring Expression Segmentation.

Fig. 3: This figure displays our model's results, showing accurate bounding boxes and segmentation masks. The figure also reveals the token selection at various layers, emphasizing how the model progressively selects the most relevant image features. The selected tokens are marked as yellow.

Qualitative Results. Fig. 3 displays our model's qualitative results and the query selection results for each layer, with the selected queries highlighted in yellow. The final output in our image demonstrates that our model can produce perfect bounding boxes and segmentation masks. Regarding query selection, in the 2nd layer, the selected queries primarily concentrate on similar instances. The selected queries become more focused in the 3rd layer, and even more so in the 4th layer, but some queries still target unrelated instances. By combining the queries from these three layers and re-ranking, we can filter out the unrelated selected queries. For instance, in the first row for 'middle man', the 2nd layer's selected queries cover all men in the image and some of the ground around them. In contrast, the 3rd layer's selected queries are more focused on the men, and the 4th layer's selected queries are entirely focused on the men, but some are on unrelated men nearby. After re-ranking, all unrelated queries about the 'middle man' are filtered out. The qualitative results demonstrate the effectiveness of our query selection mechanism.

4.2 Ablation Study

TaskSyn Network. The Table 3 shows the ablation study for our proposed components. In the first row, we evaluate our proposed TaskSyn Network, the results show our proposed TaskSyn Network can boost segmentation results by 1.45%, and bounding-box estimation results by 1.07%. These improvements are due to the decoder in Deformable DETR lacking the precise information exchange between two tasks. This experiment also demonstrates the effectiveness of our proposed TaskSyn Network in jointly learning between these two tasks.

#	Methods	REC(Pr@0.5)	RES(mIoU)
1	w/o TaskSyn Network	86.52	74.58
2	Random Queries	87.06	75.67
3	No Selection	Out of Memory	
3	Ours	**87.59**	**76.03**

Table 3: Ablation study on Refcoco val and LoA-trans-S model.

Queries Initialization Methods. The Table 3 shows the ablation study for queries selection. In #2, we show that random initialized queries can not achieve good results, because it is difficult for random initialized tokens to attend most relevant parts in multi-modality features. In #3, we use 4-th layer multi-modality features as queries for the decoder directly which reports an 'Out of Memory' issue, suggesting a computational limitation was encountered. The 'No Selection' method might imply a scenario where all multi-modality features are used without any filtering or selection process. These two results underscore the importance of efficient feature selection in managing computational resources and improving performance.

Num of Layer	REC(Pr@0.5)	RES(mIoU)
3	87.20	75.80
6	**87.59**	**76.03**
8	86.76	75.32

K	REC(Pr@0.5)	RES(mIoU)
50	86.80	75.59
100	**87.59**	**76.03**
125	86.77	75.76

Table 4: Left: Ablation study for decoder Layers on Refcoco val and LoA-trans-S model. Right: Ablation study for K selection on Refcoco val and LoA-trans-S model.

Layer of Decoder. The ablation study is presented in the Table 4 right side examines the impact of varying the number of decoder layers on two performance metrics, Pr@0.5 for bounding-box prediction accuracy and mIoU for segmentation mask prediction accuracy, in our proposed model. The optimal performance is achieved with 6 layers, yielding Pr@0.5 at 87.59 and mIoU at 76.03, which indicates a balance between model complexity and learning capability. The model with fewer decoder layers 3 shows slightly lower effectiveness, while increasing the layers to eight leads to a decrease in performance, suggesting potential overfitting or diminishing returns with added complexity.

Number of Query Selection. The ablation study is detailed in Table 4 left side elucidates the impact of TopK (K) selection on the model's final results. When K is set to 50, the model attains a Pr@0.5 score of 86.80 and an mIoU of 75.59. This performance suggests that a K value of 50 may not provide sufficient information to fully represent the mask details. Elevating K to 100 results in a noticeable enhancement in performance, elevating Pr@0.5 to 87.59 and mIoU to 76.03, indicating that this is the optimal K value for balancing detail representation and model efficiency. However, further increasing K to 125 leads to a marginal decline in performance, with Pr@0.5 decreasing to 86.77 and mIoU to 75.76. This trend suggests that beyond the optimal K value, additional information does not contribute to performance gains and could potentially introduce inefficiencies or overfitting, underscoring the necessity of a carefully calibrated K selection in our model to optimize both precision and accuracy.

4.3 Center Prediction Analysis

Centerness Score. The centerness score is introduced as a metric to evaluate the quality of center prediction in object detection tasks FCOS [32]. It measures the normalized distance from a predicted location to the center of the target object. This score provides valuable insights into how accurately a location prediction localizes the object's center, thereby aiding in assessing localization precision. To compute the centerness score, l^*, r^*, t^* and b^* corresponding to the offset between the predicted center and the center of ground-truth bounding box are used, as Equation 5. If the estimated center perfectly matches the ground truth center, the centerness is 1, while if the estimated center is on or outside of the bounding box, the centerness is 0. To evaluate the relation between centerness and final results, we split refcoco val into nine disjoint subsets according to centerness with 0.1 intervals. We then compute the performance of different models for each subset as shown in Table 5.

$$centerness = \sqrt{\frac{max(min(l^*, r^*), 0)}{max(l^*, r^*)} \times \frac{max(min(b^*, t^*), 0)}{max(b^*, t^*)}} \tag{5}$$

Referring Expression Comprehension (Pr@0.5)									
Centerness	[0.0, 0.1)	[0.2, 0.3)	[0.3, 0.4)	[0.4, 0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)	[0.9, 1.0]
TransVG [1]	30.00	29.87	40.00	42.97	55.67	63.74	80.54	81.77	89.64
SeqTR [46]	**30.63**	28.57	**46.15**	46.09	50.74	63.51	79.97	82.46	90.38
LoA-Trans	10.95	**37.05**	45.14	**53.30**	**75.81**	**91.78**	**97.52**	**99.44**	**99.93**
LoA-Trans*	34.40	52.94	59.42	66.06	85.25	94.08	97.89	99.44	99.93
Referring Expression Segmentation (mIoU)									
Centerness	[0.0, 0.1)	[0.2, 0.3)	[0.3, 0.4)	[0.4, 0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)	[0.9, 1.0]
SeqTR [46]	29.45	25.83	41.13	35.16	44.82	53.47	60.45	64.92	73.97
LAVT [37]	**30.62**	33.48	44.22	45.22	50.33	58.77	67.61	73.94	83.04
LoA-Trans	15.95	**42.93**	**48.71**	**54.60**	**63.86**	**70.93**	**78.11**	**84.96**	**89.03**
LoA-Trans*	18.65	45.35	50.19	56.68	64.64	71.67	78.42	85.08	89.05
#samples	904	136	175	221	339	608	1089	2657	4609

Table 5: Upper: The relation between centerness score and Pr@0.5 in referring expression comprehension task. Bottom: The relation between centerness score and mIoU in referring expression segmentation task. LoA-Trans with $*$ means the results for replacing the prediction center with the ground-truth center (i.e., what if the center prediction is perfect). #samples means the number of samples falls in this centerness range. This experiment is done on Refcoco val dataset and LoA-trans-B model.

Centerness vs. Box IoU. Our investigation into the relationship between Centerness [32] and Box IoU, as shown in Table 5 upper, revealed some interesting patterns. Centerness, is the metric indicating how accurately the bounding box has been placed over the object. As expected, when the targeting quality is high(centerness score is high), the bounding box quality tends to be higher. Interestingly, our data also indicated instances where even with less optimal targeting quality, as evidenced by lower Centerness scores, the bounding boxes still had high Pr@0.5 scores. This suggests that in some cases, our decoder can compensate for objects not being perfectly targeted and still predict their bounding boxes accurately. More details can be found in supplementary materials Fig.2.

Centerness vs. Segmentation IoU. Our analysis of Centerness about Segmentation IoU, as shown in Table 5 bottom shed light on how targeting quality correlates with segmentation accuracy. We found a trend that mirrors our expectations: segments with a high Centerness score frequently aligned with higher Segmentation IoU scores, indicating more accurate segmentations. This was particularly true for segments that were well-defined and distinct. Yet, there were interesting cases where segments with lower Centerness scores still achieved a high Segmentation IoU. This points to the possibility that our segmentation approach can tolerate some degree of off-centering without a substantial loss in accuracy. More detail can be found in supplementary materials Fig.2.

Ground-Truth Center vs. Estimated Center. To evaluate the effectiveness center prompt, we conducted an experiment where we replaced the estimated center with the ground-truth center during validation to investigate the potential upper-bound performance. This approach allows us to assess the ideal scenario where the center is precisely known. The results are provided in Table 5 LoA-Trans with $*$. In assessing the performance disparity between Ground-Truth Center and Estimated Center across the metrics of Pr@0.5 in bounding-box prediction accuracy and mIoU in segmentation mask accuracy, our analysis reveals notable differences. On average, the results of GroundTruth Center surpass the results of Estimated Center both in bounding box prediction and segmentation mask estimation. These results demonstrate that accurate center estimation plays a critical role in bounding box prediction and segmentation mask estimation. The precision in center estimation directly influences the accuracy of locating objects within an image, which is a fundamental aspect of effective bounding box prediction and segmentation mask estimation. Consequently, the effectiveness of the Center Estimated Module not only enhances object detection but also ensures more accurate and reliable segmentation, underscoring its importance in multi-task learning. More references can be found in supplementary materials Fig. 3.

Comparsion with Other Methods. In our comparative analysis of model performance across varying centerness intervals as shown in Table 5, a clear trend emerges, underscoring the integral role of centerness in enhancing model precision in referring expression comprehension and segmentation tasks. Notably, LoA-Trans demonstrates superior performance, particularly in higher centerness ranges, outshining competitors like TransVG [1] in REC, LAVT [37] in RES, and both REC and RES in SeqTR [46]. In detail, if centerness is above 0.5, our method significantly improves the performance compared to the other methods. Only if centerness is significantly small (e.g., less than 0.3) our performance is inferior, it's still acceptable, however. If centerness is very very small (less than 0.1) our performance is significantly lower than the others. This analysis not only spotlights the pivotal role of centerness in visual comprehension tasks but also sets the stage for further introspection into the mechanisms driving LoA-Trans's exceptional performance, paving the way for future advancements in model development and refinement.

Corner Cases. As mentioned above, there are two special cases in our model: (a) the estimated centerness is low, but the model can predict bounding-box and segmentation mask correctly, as shown in Fig.4 left side. It is noticeable that the estimated center consistently appears close to the ground truth bounding box. This slight displacement could be attributed to the fact that we utilize the smallest feature maps for center pre-

diction. Thanks to the proposed query selection mechanism, the model can focus on the most relevant features related to target objects, improving the performance of detection and segmentation. (b) In the cases depicted on the right side of Fig. 4, the estimated centerness scores are high, indicating confidence in object localization. However, the IoU for both the predicted bounding box and segmentation mask is low. In the first two rows, although the model correctly identifies the objects, the size estimation is inaccurate due to complex scenarios, leading to lower IoU values. In the last rows, the model locates the center of the ground truth bounding box, but the bounding box center does not correspond to the target objects; instead, it is near a similar instance. This misalignment results in low IoU values despite the high centerness scores.

Fig. 4: Some corner cases in our proposed methods. The green triangle indicates the estimated center. Left: low centerness score but high detection and segmentation IoU. Right: high centerness score but low detection and segmentation IoU.

5 Conclusion

In our work, we presented LoA-Trans, a cutting-edge method focusing on joint learning of Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES). Our approach innovatively commences by introducing a center prompt as location guide for REC and RES. And query selection mechanism is used to generate queries for the decoder, enabling the capture of the complex interrelations within the multi-modality feature maps. Furthermore, the introduction of the TaskSyn network is a pivotal aspect of LoA-Trans, designed to enable effective information exchange between REC and RES. This unique mechanism ensures that one task benefit the other, creating a synergistic effect that enhances both comprehension and segmentation capabilities. Our extensive evaluations and comparisons have demonstrated that LoA-Trans surpasses existing methods in performance, setting new state-of-the-art benchmarks.

Acknowledge

This work is partly supported by JSPS KAKENHI Grant Number JP23K24876 and JST ASPIRE Program Grant Number JPMJAP2303.

References

1. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: ICCV (2021)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV (2021)
4. Feng, G., Hu, Z., Zhang, L., Lu, H.: Encoder fusion network with co-attention embedding for referring image segmentation. In: CVPR (2021)
5. Girshick, R.: Fast r-cnn. In: ICCV (2015)
6. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. IEEE TPAMI (2019)
7. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR (2017)
8. Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: CVPR (2020)
9. Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: CVPR (2020)
10. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: ICCV (2021)
11. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
12. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. In: NeurIPS (2021)
13. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: CVPR (2020)
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
16. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR (2023)
17. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: ICCV (2017)
18. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: ICCV (2019)
19. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: CVPR (2023)
20. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. In: CVPR (2019)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
24. Luo, G., Zhou, Y., Ji, R., Sun, X., Su, J., Lin, C.W., Tian, Q.: Cascade grouped attention network for referring expression segmentation. In: ACM MM (2020)
25. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: CVPR (2020)

26. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
27. Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: ECCV (2018)
28. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016)
29. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (2019)
30. Su, W., Miao, P., Dou, H., Wang, G., Qiao, L., Li, Z., Li, X.: Language adaptive weight generation for multi-task visual grounding. In: CVPR (2023)
31. Sun, M., Xiao, J., Lim, E.G.: Iterative shrinking for referring expression grounding using deep reinforcement learning. In: CVPR (2021)
32. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV (2019)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
34. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: CVPR (2022)
35. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: ICCV (2019)
36. Yang, S., Xia, M., Li, G., Zhou, H.Y., Yu, Y.: Bottom-up shift and reasoning for referring image segmentation. In: CVPR (2021)
37. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR (2022)
38. Yang, Z., Chen, T., Wang, L., Luo, J.: Improving one-stage visual grounding by recursive sub-query construction. In: ECCV (2020)
39. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: ICCV (2019)
40. Ye, J., Tian, J., Yan, M., Yang, X., Wang, X., Zhang, J., He, L., Lin, X.: Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In: CVPR (2022)
41. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR (2019)
42. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR (2018)
43. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016)
44. Zhang, H., Niu, Y., Chang, S.F.: Grounding referring expressions in images by variational context. In: CVPR (2018)
45. Zhou, Y., Ji, R., Luo, G., Sun, X., Su, J., Ding, X., Lin, C.W., Tian, Q.: A real-time global inference network for one-stage referring expression comprehension. IEEE TNNLS (2021)
46. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: ECCV (2022)
47. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)