VLFormer: Visual-Linguistic Transformer for Referring Image Segmentation

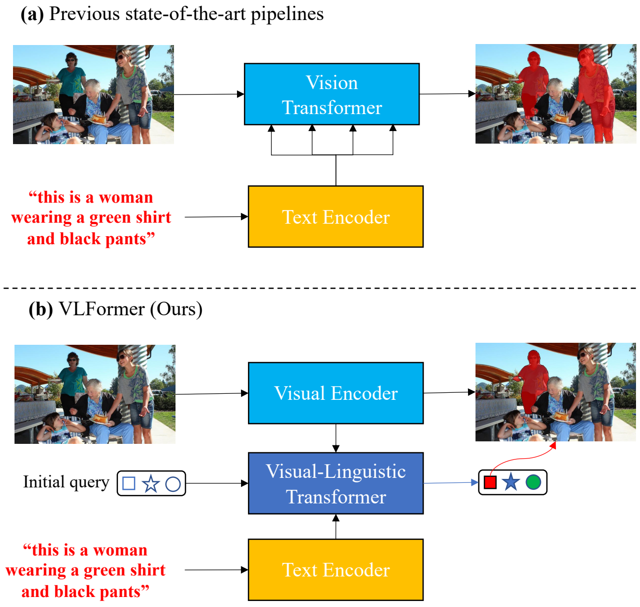
Submission # 5363

Abstract

The referring image segmentation task aims to segment a referred object from an image using a natural language expression. The query expression in referring image segmentation typically describes the relationship between the target object and others. Therefore, several objects may appear in the expression, and the model must carefully understand the language expression and select the correct object that the expression refers to. In this work, we introduce a unified and simple query-based framework named VLFormer. Concretely, we use a small set of object queries to represent candidate objects and design a mechanism to generate the fine-grained object queries by utilizing language and multi-scale vision information. More specifically, we propose a Visual-Linguistic Transformer Block, which produces a richer representation of the objects by associating visual and linguistic features with the object queries effectively and simultaneously. At the same time, we leverage the ability to extract linguistic features from CLIP, which has a great potential for compatibility with visual information. Without bells and whistles, our proposed method significantly outperforms the previous state-of-the-art methods by large margins on three referring image segmentation datasets: RefCOCO, RefCOCO+, and G-Ref.

1 Introduction

Referring image segmentation aims to predict a pixel-wise mask of the referred object given an image and a natural language expression. This task can be potentially used in a wide range of applications, including human-object interaction and image editing. Unlike traditional visual segmentation tasks (such as semantic segmentation (He et al. 2019) and instance segmentation (He et al. 2017)) that require a fixed number of categories, referring image segmentation has to deal with a broader amount of vocabularies and syntax diversities of human languages. In this task, the target object is mentioned with various forms of expression, such as words, phrases, or complex sentences presenting the concepts of actions, positions, objects, etc. Hence, the most challenging part of this task is to understand the expression and highlight the regions that are relevant to that expression.



(a) Previous state-of-the-art pipelines
Vision Transformer
Text Encoder
"this is a woman wearing a green shirt and black pants"
(b) VLFormer (Ours)
Visual Encoder
Initial query
Visual-Linguistic Transformer
Text Encoder
"this is a woman wearing a green shirt and black pants"

Figure 1: Comparison of referring image segmentation pipelines. (a) The previous state-of-the-art approach (i.e., LAVT (Yang et al. 2022)) integrate linguistic features into visual features of a vision transformer model to benefit the jointly exploiting vision-language cues. (b) We propose to leverage a Transformer-based module to associate both visual and linguistic information with a set of object queries, aiming to gradually update the representation of these object queries. The final object queries then produce the mask prediction.

Over the past few years, the referring image segmentation task has grown rapidly. Early approaches (Liu et al. 2017; Margffoy-Tuay et al. 2018; Li et al. 2018) leverage a fusion module between visual and linguistic features and followed by a cross-modal decoder to generate masks of the referred object. Concretely, the fusion modules includes recurrent interaction (Liu et al. 2017; Li et al. 2018), cross-modal attention (Shi et al. 2018; Chen et al. 2019a), language-guided modeling (Huang et al. 2020; Hui et al. 2020), etc. Recently, Transformer (Vaswani et al. 2017) shows a significant improvement in performance of cross-modal alignments (Ding et al. 2021; Yang et al. 2022; Botach, Zheltonozhskii, and Baskin 2022) (illustrated in Fig. 1).

In most previous transformer-based works in computer vision (Carion et al. 2020; Wang et al. 2021; Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022), a set of queries is used to represent the class or instance features for detection and segmentation. In referring segmentation task, some works (Ding et al. 2021; Wu et al. 2022) generate the query vectors from language features using vision-guided attention or directly. Then these queries are updated in a Transformer decoder using visual features only. It can lead to a problem that the linguistic information can vanish in the object query features after several Transformer decoder layers. To address this issue, a potential solution is to exploit a multi-modal Transformer for simultaneously aggregating visual and linguistic features during the transformer decoder.

Besides, on such referring segmentation problems, the text query usually contains information about the category, position on the image, and appearance of the object. In some cases, the related position between the referred object and others is mentioned. CLIP (Radford et al. 2021) models are learned from a wide range of visual concepts followed by the natural language. Hence, these models can capture the information related to the visual concepts better.

Hence, we propose a Visual-Linguistic Transformers (VLFormer) approach to leverage a set of queries that understand both visual and linguistic features to represent potential objects. First, the CLIP Text Encoder is utilized for linguistic features from natural language expression. The linguistic features are used to improve the vision-language fusion and enhance the representation of object queries through the Transformer-based module. Second, a Visual-Linguistic Transformer, which contains several Visual-Linguistic Transformer Block modules, is designed for constructing fine-grained object features using linguistic and multi-scale visual information. Each Visual-Linguistic Transformer Block (VLB) utilizes the cross-attention modules from linguistic features and visual features to object queries, then generates more informative object queries. Figure 1 highlights the difference between our proposed work and the state-of-the-art method.

To evaluate the effectiveness of our framework, we conduct experiments on several widely used datasets for referring image segmentation. Our VLFormer achieves 74.67%, 64.80%, and 66.77% IoU on the validation sets of RefCOCO, RefCOCO+, and G-Ref, respectively. Our method improves the state-of-the-art for these datasets by large margins of 1.94%, 2.66%, and 5.53%, respectively.

To summarize, our main contributions are listed as follows:

- We propose VLFormer, a query-based referring image segmentation framework that efficiently interacts among multi-modal features via a Transformer-based module.
- We design a Visual-Linguistic Transformer Block, which associates visual and linguistic features with the object queries to produce a more fine-grained representation of object queries.
- We suggest utilizing the cross-modal knowledge of the CLIP model for achieving better vision-language fusion.
- We achieve new state-of-the-art results on three datasets for referring image segmentation: RefCOCO, RefCOCO+, and G-Ref.

2 Related Work

Referring Image Segmentation. Referring image segmentation was first introduced by (Hu, Rohrbach, and Darrell 2016) to segment a target object or stuff in an image given a natural language expression. Early works proposed extracting visual and linguistic features independently from CNN and RNN, respectively, then concatenating them to gather multi-modal features and generating the final segmentation results by a decoder. Utilizing the successful of instance segmentation, Yu et al. (Yu et al. 2018) proposed a two-stage method that first generates a set of candidate instances, then leverages the linguistic features as guidance for selecting the referred object from that set. Recent works (Ding et al. 2021; Yang et al. 2022) design a Transformer-based multi-modal encoder aiming to fuse the visual and linguistic features. It captures the interaction between vision and language information in the early phase. In (Ding et al. 2021), the input image and language expression adaptively produce a set of queries, then these queries are used to generate the referred object mask through a Transformer decoder. Different from previous methods, CRIS (Wang et al. 2022) leverages the knowledge of the CLIP (Radford et al. 2021) and text-to-pixel contrastive learning to improve the compatibility of multi-modal information and boost the ability of cross-modal matching.

Query-based Architecture. Recently, query-based architectures have achieved marvelous success in most computer vision tasks such as object detection (Carion et al. 2020), object tracking (Meinhardt et al. 2022), and image segmentation (Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022). The main component in these architectures is a Transformer model with a set of object queries.

DETR (Carion et al. 2020) introduced the new query-based paradigm for object detection. The image features were encoded by a transformer encoder, then these encoder outputs were used to transform a set of object queries into the output embedding by the decoder. An extension of this method for Video Instance Segmentation (VIS) is presented in VisTR (Wang et al. 2021). Wang et al. proposed a new strategy of instance sequence matching and segmentation to supervise and segment instances at the sequence level simultaneously (Wang et al. 2021). Cheng et al. introduced MaskFormer (Cheng, Schwing, and Kirillov 2021) as a new mask classification instead of the per-pixel classification for segmentation tasks. MaskFormer shows the simplicity and convenience of the general segmentation paradigm by using a set of queries as semantic regions. Mask2Former (Cheng et al. 2022) extends the meta-architecture of MaskFormer with a new Masked-attention transformer decoder and utilizes high-resolution features to handle objects on multiple scales.

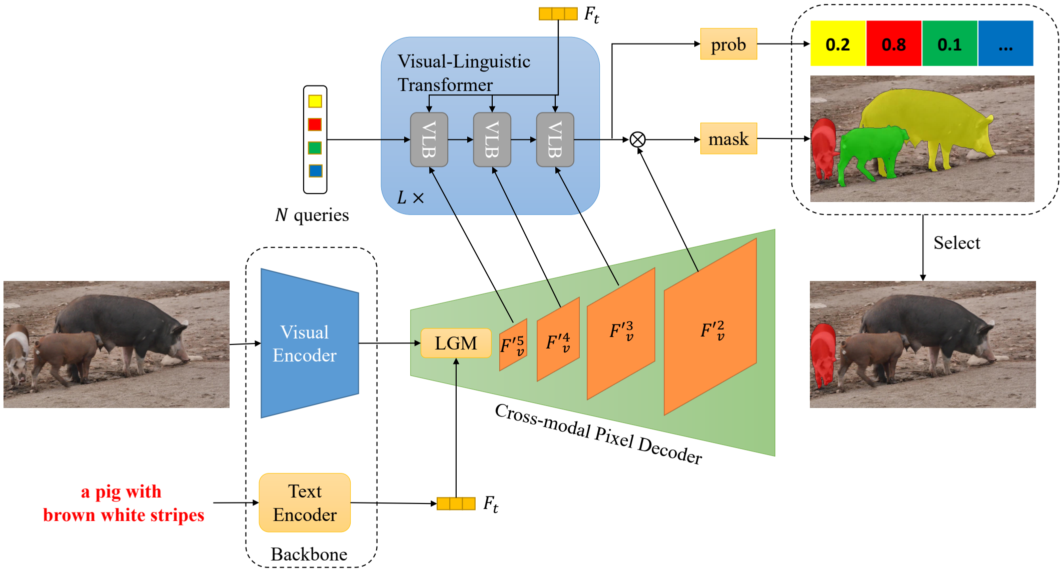


Figure 2: The overview of our method VLFormer. It contains three major components: Backbone (includes Visual Encoder and Text Encoder), Cross-modal Pixel Decoder, and Visual-Linguistic Transformer. First, the model extracts visual features and linguistic features from a given image and language expression by the backbone. The visual features are then processed by a Cross-modal Pixel Decoder to generate fine-grained language-guided visual features. N object queries are gradually updated by the Visual-Linguistic Transformer and then are multiplied with per-pixel embedding to output N potential objects. The final object is selected based on the probability generated by the object queries.

3 Method

In this section, we first present the overview of our proposed VLFormer. We then describe how we extract the visual and linguistic features from image and language expression, respectively. We further introduce the upgraded Cross-modal Pixel Decoder and our proposed Visual-Linguistic Transformer. Finally, we provide the details of the instance matching and loss function used in the training phase.

3.1 Overview

Given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a referring expression T with L words, where H and W are the height and the width of the image, respectively. We propose an efficient and simple end-to-end framework named VLFormer for referring image segmentation, as shown in Figure 2. It consists of three main components: Backbone, Cross-modal Pixel Decoder, and Visual-Linguistic Transformer.

First, a backbone includes Visual Encoder and Text Encoder to extract low-resolution feature maps from an image and linguistic features from the corresponding referring expression, respectively. Next, a Cross-modal Pixel Decoder contains a Language Guidance Module followed by a Pixel Decoder adopted from Mask2Former (Cheng et al. 2022). It incorporates coarse features from the visual features with the linguistic features by a Language Guidance Module and upsamples these features to generate the fine-grained language-guided per-pixel embeddings. Then, Visual-Linguistic Transformer modules operate on linguistics and multi-scale visual features from the Cross-modal Pixel Decoder to enrich the object queries. Finally, the binary mask predictions are generated by multiplying the per-pixel embeddings and object queries and followed by an Instance Matching process in the training phase. During inference, we select the query with the highest confidence score as the target object for the final output.

3.2 Feature Extraction

As illustrated in Figure 2, the input of our framework consists of a video V and a referring expression with L words.

Visual Encoder For an image, $I \in \mathbb{R}^{H\times W\times 3}$, we extract the multi-scale feature maps by a visual encoder. The visual encoder, e.g., ResNet, usually processes an image with five different stages that correspond to different scales: $1/2$, $1/4$, $1/8$, $1/16$, and $1/32$.

In our model, we utilize the four last stages and transform their feature dimension into a common dimension C by a Multilayer Perceptron (MLP) module for convenience. We obtain the multi-scale feature maps $\{F_v^j\} \in \mathbb{R}^{\frac{H}{2^j}\times\frac{W}{2^j}\times C}$ for each stage $j \in (2, 3, 4, 5)$, respectively. Note that the H and W are the height and width of the original image.

Text Encoder. We adopt the text encoder of the CLIP (Radford et al. 2021) model to extract the linguistic features $F_t \in \mathbb{R}^{L\times C'}$ from the L-word expression. The text encoder of the CLIP model is a modified Transformer module. For convenience in later stages (e.g., Transformer decoder), we compress the dimension C' to C by a MLP module so that the dimension of linguistic features is now the same as the visual features. We also add a sinusoidal 1D positional encoding $e_{pos} \in \mathbb{R}^{L\times C}$ to the linguistic features to store the positional information in these features.

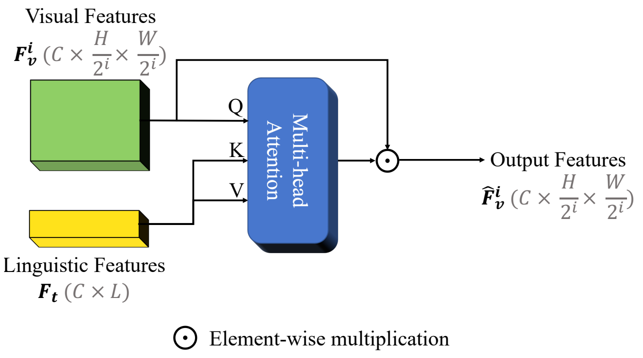

Visual Features
$F_v^i \, (C \times \frac{H}{2^i} \times \frac{W}{2^i})$
Q
K
V
Multi-head Attention
Output Features
$\hat{F}_v^i \, (C \times \frac{H}{2^i} \times \frac{W}{2^i})$
Linguistic Features
$F_t \, (C \times L)$
⊙ Element-wise multiplication

Figure 3: The implementation of Language Guidance Module. Visual features F_v^i and linguistic features F_t is used as input to output the language-guided visual features $\hat{F}_v^i$.

3.3 Cross-modal Pixel Decoder

The multi-scale visual features that are extracted by the visual encoder then are incorporated with the linguistics features, aiming to enrich the information related to the referred object in the image. Then will be gradually decoded again into the fine-grained features.

Language Guidance Module Firstly, the visual features are fused with the linguistic features F_t, extracted by the text encoder to perform an early interaction between visual and linguistic features by a Language Guidance Module (LGM) for highlighting regions that are matched with the referring expressions.

$$\hat{F}_v^i = F_v^i \odot MSA(F_v^i W^Q, F_t W^K, F_t W^V), \quad (1)$$

where MSA(q, k, v) is the multi-head attention layer and W^Q, W^K, $W_V \in \mathbb{R}^{C \times d_{head}}$ are learnable parameters. This multi-head attention layer is used as a compatibility weight between the visual features and linguistic features. Then, this weight is multiplied with the visual feature to focus on region highly related to referring expressions. Figure 3 shows the LGM implementation.

Pixel Decoder The pixel decoder gradually upsamples the visual features $\hat{F}_v^i$ to generate the fine-grained language-guided visual features F'^{i}_v. The Pixel Decoder is adopted from Mask2Former (Cheng et al. 2022) with a multi-scale deformable attention module to produce a fine-grained feature pyramid.

3.4 Visual-Linguistic Transformer

In order to aggregate the linguistic information and multi-scale visual features into object queries, a Visual-Linguistic Transformer is designed with three Visual-Linguistic Transformer Block to leverage different scales of visual features.

Visual-Linguistic Transformer Block

We propose a Visual-Linguistic Transformer Block (VLB) to model the interaction between queries and visual features and between queries and linguistic features simultaneously. As illustrated in Figure 4, it uses the visual features, linguistic features of the referring expressions, and the object queries features as its input.

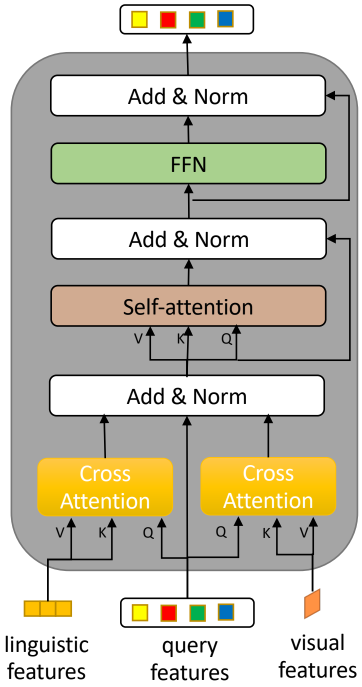


Figure 4: The architecture of Visual-Linguistic Transformer Block (VLB). The module takes query features, linguistic features F_t and visual features F'^{i}_v as input, and generates a better representation of query features.

First, the object queries interact independently with linguistic and visual features in a cross-attention way, where object queries play roles as query, both linguistic and visual features are key and value. And then, they will be summed up and fed into a multi-head self-attention module to update the object queries features and gather the contextual information of both linguistic and visual features. Finally, a FFN module is applied to these features to get the final object query features.

Multi-scale features Multi-scale features can help the model improve its performance, especially for various object sizes. By utilizing both low-resolution features and high-resolution features in the Transformer decoder layers, the object features can adapt to the size-changing and diversity of the object's shapes. Specifically, we utilize the features F'^{3}_v, F'^{4}_v, and F'^{5}_v produced by the cross-modal pixel decoder with resolutions 1/8, 1/16, and 1/32 of the original image, respectively. As illustrated in Figure 2, in one Transformer decoder layer, there are 3 Visual-Linguistic Transformer blocks. In the first block, the object queries will be updated by the linguistic features and the visual features F'^{5}_v. In the second VLB, the updated object queries will be re-updated by the linguistic and the visual features F'_{v4}. The last one updates the object queries in the same way but uses the visual feature F'^{3}_v instead.

By repeating these operations multiple times, the final object queries can be adapted with multiple-scale features and can also handle the variety of object sizes in an image.

Method	Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	testA	testB	val	testA	testB	val	test
MAttNet (Yu et al. 2018)	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree (Liu et al. 2019)	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA (Ye et al. 2019)	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2Seg (Chen et al. 2019b)	ResNet-101	58.90	61.77	53.81	-	-	-	46.37	46.95
CMPC (Huang et al. 2020)	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-
EFNet (Feng et al. 2021)	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
VLT (Ding et al. 2021)	DarkNet-53	65.65	68.29	62.73	55.5	59.2	49.36	52.99	56.65
ReSTR (Kim et al. 2022)	ViT-B-16	67.22	69.3	64.45	55.78	60.44	48.27	54.48	-
CRIS (Wang et al. 2022)	ResNet-101	70.47	73.18	66.1	62.27	68.08	53.68	59.87	60.36
LAVT (Yang et al. 2022)	Swin-B	72.73	75.82	68.79	62.14	68.38	55.1	61.24	62.09
VLFormer(**Ours**)	ResNet-50	73.92	76.03	**70.86**	64.02	69.74	55.04	65.69	65.90
VLFormer(**Ours**)	ResNet-101	**74.67**	**76.8**	70.42	**64.80**	**70.33**	**56.33**	**66.77**	**66.52**

Table 1: Comparisons with the state-of-the-art approaches on three benchmarks. We report the results of our method with various visual backbones. IoU is used as the main metric, and "-" shows that the result is not available. The best performance is marked in boldface.

3.5 Instance Matching and Loss

Our approach is to generate a small set of N predictions, and the best one will be selected as the final object. In the training phase, therefore, we use the instance matching strategy inspired by (Cheng et al. 2022; Cheng, Schwing, and Kirillov 2021) to supervise candidate instances. Let us denote the prediction set as $\hat{y} = \{\hat{y_i}\}_{i=1}^{N}$, and the prediction for the i-th instance is represented by:

$$\hat{y}_i = \{\hat{p}_i, \hat{s}_i\}. \tag{2}$$

For the i-th candidate instance, $\hat{p}_i \in \mathbb{R}^1$ is a probability that this instance corresponds to the referred object. Meanwhile, $\hat{s}_i \in \mathbb{R}^{H \times W}$ is the segmentation mask that we predict.

Since there is only one referred object, the ground-truth object is represented as $y \in \mathbb{R}^{H \times W}$. To train the network, we first find the best prediction i-th from N candidates via minimizing the matching cost:

$$\begin{aligned}\mathcal{L}_{match}(y, \hat{y}_i) = \gamma_{cls}\mathcal{L}_{cls}(\hat{p}_i, 1) + \gamma_{mask}\mathcal{L}_{mask}(\hat{s}_i, y) \\ +\gamma_{dice}\mathcal{L}_{dice}(\hat{s}_i, y).\end{aligned} \tag{3}$$

The matching cost is computed based on three loss functions. First, $\mathcal{L}_{cls}$ represents the loss function for the probability a query is the referred object, and we use the Cross-Entropy loss in this work. Second, $\mathcal{L}_{mask}$ is the cross entropy loss that is designed to supervise the mask prediction. Finally, the $\mathcal{L}_{dice}$ is added to improve the dice score, which is quite similar to the IoU metric. $\gamma_{cls}, \gamma_{mask}$ and γ_{dice} are the coefficients of the three corresponding losses.

Our goal is to minimize the $\mathcal{L}_{match}$ of one query and maximize the probability of other queries $\hat{p}_j (j \neq i)$ approximate zero (it means these queries do not represent for the referred object). Therefore, our loss function is described as follows:

$$\mathcal{L}(y, \hat{y}, i) = \mathcal{L}_{match}(y, \hat{y}_i) + \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_{cls}\mathcal{L}_{cls}(\hat{p}_j, 0). \tag{4}$$

4 Experiments

4.1 Datasets

We evaluate our VLFormer on three commonly used datasets for referring image segmentation: RefCOCO, RefCOCO+ (Kazemzadeh et al. 2014) and G-Ref (Mao et al. 2016).

RefCOCO & RefCOCO+ were collected using the two-player ReferitGame. In this game, the first player is given an image with a segmented object and asked to write a language expression to describe the target object. The second one is shown only the image and the referring expression and asks to choose the corresponding object. RefCOCO has 142,209 expressions for 50,000 objects in 19,994 images, and RefCOCO+ consists of 141,564 expressions for 49,856 objects in 19,992 images. Some kinds of words, e.g., words about absolute locations, are not used in the RefCOCO+ dataset. Therefore, it is considered to be more challenging than the RefCOCO dataset. The expressions in RefCOCO and RefCOCO+ are very concise, contains 3.5 words on average, and tend to have more objects of the same category per image (3.9 on average).

G-Ref is another commonly widely used dataset. It contains 104,560 expressions referring to 54,822 objects belonging to 26,711 images. Compared to RefCOCO and RefCOCO+, the average sentence length in G-Ref dataset is longer (8.4 words on average), and G-Ref has a richer word usage. However, G-Ref has fewer objects of the same category per image than RefCOCO and RefCOCO+ (1.6 on average).

4.2 Implementation Details

Experimental Settings. We implement our method in PyTorch and use the CLIP Text Encoder implementation from HuggingFace's Transformer library. The Text Encoder is frozen during the training stage. We choose the number of Visual-Linguistic Transformer layers is $L = 2$, which contains six VLB layers in total. The common feature dimen-

	Precision@0.5	Precision@0.6	Precision@0.7	Precision@0.8	Precision@0.9	IoU
(a) Number of Visual-Linguistic Transformer layers (L)						
0	79.67	76.60	71.77	61.89	33.36	70.21
1	83.27	80.56	76.43	66.81	37.78	73.45
2	**83.82**	**81.18**	**76.84**	**67.48**	**37.89**	**73.92**
(b) Text Encoder model						
CLIP	**83.82**	**81.18**	**76.84**	**67.48**	**37.89**	**73.92**
BERT	79.85	77.54	73.80	65.23	36.62	70.73
(c) Language Guidance Module (LGM)						
With LGM	**83.82**	**81.18**	**76.84**	**67.48**	**37.89**	**73.92**
W/o LGM	73.93	71.26	67.42	58.81	33.14	65.19
(d) Number of object queries (N)						
1	82.47	78.85	73.34	62.90	33.31	72.59
3	83.19	80.31	75.60	65.97	36.41	73.18
5	**83.82**	**81.18**	**76.84**	**67.48**	**37.89**	**73.92**
8	82.12	79.44	75.03	65.89	36.28	72.47
10	82.77	80.29	75.98	66.52	36.27	73.04

Table 2: Ablation Study on RefCOCO. The experiments are based on ResNet-50 visual backbone and conducted on the validation split of RefCOCO. W/o LGM indicates that LGM is not used in the Cross-modal Pixel Decoder. The best performance is marked in boldface.

sion C is set to 256. In the RefCOCO, RefCOCO+, and G-Ref dataset, we train the network for 100K iterations using the AdamW optimizer with the initial learning rate 0.0001. Then a factor of 0.1 decreases the learning rate at the 70K-th iteration. The network is trained with a small batch size of 8 on 2 NVIDIA RTX 2080Ti with 12GB GPU VRAM.

Metrics. There are two metrics we use for experiments, namely, IoU and Precision@X. The IoU score shows the quality of the prediction overlapped with the ground-truth, which demonstrates the overall performance of the approach. The Precision@X reports the successful referring rate at the threshold X of the IoU score, which focuses on the referring ability of the method.

4.3 Comparison with State-of-the-art

We compare our proposed method with several state-of-the-art methods on three common datasets for referring image segmentation. As illustrated in Table 1, our method suppresses other methods on each split of all datasets with large margins. The experiments demonstrate our method can surpass the state-of-the-art in several split of datasets even though we leverage the ResNet-50 visual backbone (He et al. 2016).

On the RefCOCO dataset, our model significant outperforms the state-of-the-art LAVT (Yang et al. 2022) by 1.94%, 0.98% and 1.63% on three splits, respectively, even our ResNet-101 visual backbone instead of Swin Transformer-Base. Similarly, VLFormer achieves an impressive performance improvement of around 2% than several state-of-the-art methods on each split of the more challenging RefCOCO+ dataset.

Besides, in the most challenging dataset G-Ref, which has longer and more complex expressions, our proposed VLFormer outperforms the previous state-of-the-art methods with wide margins of 4.97% and 4.43% on the validation and test subsets, respectively. As shown in Table 1, the results demonstrate that our proposed approach has the powerful ability to understand long and complex sentences.

4.4 Ablation Study

In this section, we perform extensive ablation studies on the validation set of RefCOCO to study the effect of core components in our model. All models are based on ResNet-50 visual backbone. The details analysis is as follows.

Visual-Linguistic Transformer Table 2(a) reports the performance of our framework in various number of Visual-Linguistic Transformer layers. Without Visual-Linguistic Transformer Block(VLB) ($L = 0$), the model uses directly initialized object queries to associate with the output from Cross-modal Pixel Decoder to generate the object segmentation. In this case, the object queries play a role as the fully-connected layer to perform a per-pixel segmentation since the object queries are not updated by either vision or language information. As shown in Table 2, removing all VLB modules leads to a drop of 3.71% in IoU metric and a drop of 4% to 6% in precision across five thresholds. The setting of $L = 1$ immediately improves the performance with an increase of 3.24% in IoU. Then our VLFormer consistently increases the results with more Visual-Linguistic Transformer layers. These results illustrate the effectiveness of aggregating the visual and linguistic features with object queries via our proposed VLB module to enrich the object representation ability. We choose $L = 2$ as the default in our framework to keep it simple and efficient.

Text Encoder. Table 2(b) shows that using BERT (Devlin et al. 2019) to extract linguistic features instead of CLIP Text

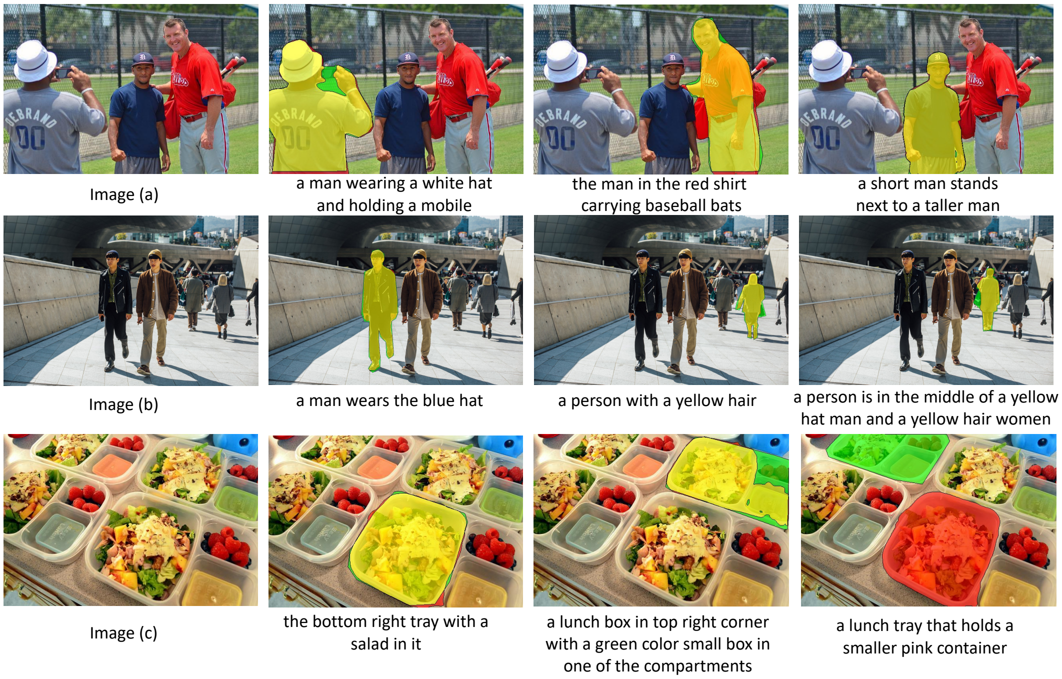


Image (a)
a man wearing a white hat and holding a mobile
the man in the red shirt carrying baseball bats
a short man stands next to a taller man
Image (b)
a man wears the blue hat
a person with a yellow hair
a person is in the middle of a yellow hat man and a yellow hair women
Image (c)
the bottom right tray with a salad in it
a lunch box in top right corner with a green color small box in one of the compartments
a lunch tray that holds a smaller pink container

Figure 5: The visualization of referring image segmentation results. From left to right: the input image, our overlaid results with the corresponding expressions. Overlaid color legend: green and red indicate the groundtruth and our segmentation results, respectively; while yellow highlights the intersection between the ground truth and our results.

Encoder leads to a drop of 3.19% in IoU. In addition, the precision drops by 2% to 4% in all the thresholds from 0.5 to 0.9. These results demonstrate the benefit of exploiting the ability to interact with visual features using linguistic features extracted from the CLIP Text Encoder model.

Language Guidance Module (LGM). In Table 2(c), we compare the standard Pixel Decoder and the Cross-modal Pixel Decoder, which leverages the Language Guidance Module to re-weight multi-scale visual features by the linguistic features. The results show that the Language Guidance Module provides more accurate segmentation by an improvement of around 8% in IoU score and $4-10\%$ of precision in several thresholds. It indicates that guiding the visual features with the expression is essential in the referring segmentation.

Number of Queries. To demonstrate the effectiveness of the query number N, we show our VLFormer's performance with several numbers of queries in Table 2(d). Benefitting from the Visual-Linguistic Transformer Block design, all the initial object queries are learned to incorporate both linguistic and visual features robustly to find the referred object. More queries can help the model make judgments among potential instances, which could handle the similarity of objects in complicated scenes. The performance is highest at $N = 5$ and begins slightly decrease when the number of queries gets larger. When $N = 1$, the referring expression may be complicated and it causes confusion to the referred target object.

4.5 Qualitative Results

We visualize the referring image segmentation results from our method in Figure 5. To illustrate the impressive ability of our method, we show the predicting results of some examples with different expressions. Image (a) shows that our method can handle the expression containing the color information. The last image in the first row even demonstrates the capability to segment objects with attributes about the relative height, i.e., short, tall. In (b), we can see that our network can understand the color, i.e., blue, yellow and related stuff, i.e., hat, hair, and also identifies the referred person who locates in the middle of the two described objects. In (c), VLFormer correctly identifies the correct objects in the first two expressions. However, the third expression regarding the "smaller pink container" is confusing since there is no obvious pink container in the image. Therefore, VLFormer picks a pinkish object. This incorrect segmentation illustrates the failure case in our proposed work.

5 Conclusion

In this paper, we propose VLFormer framework for referring image segmentation. We design a Visual-Linguistic Transformer Block (VLB) to enrich the object queries by associating them with visual and linguistic features through a transformer-based module. We also successfully utilize Text Encoder from the CLIP model to generate a better representation of linguistic features for cross-modal fusion. Experimental results have demonstrated that our proposed method outperforms previous state-of-the-art methods with large margins on three widely used datasets.

References

Botach, A.; Zheltonozhskii, E.; and Baskin, C. 2022. End-to-End Referring Video Object Segmentation With Multimodal Transformers. 4985–4995.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, 213–229. Cham: Springer International Publishing. ISBN 978-3-030-58452-8.

Chen, D.-J.; Jia, S.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019a. See-Through-Text Grouping for Referring Image Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7453–7462. ISSN: 2380-7504.

Chen, Y.-W.; Tsai, Y.-H.; Wang, T.; Lin, Y.-Y.; and Yang, M.-H. 2019b. Referring Expression Object Segmentation with Caption-Aware Consistency. ArXiv:1910.04748 [cs].

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-Attention Mask Transformer for Universal Image Segmentation. 1290–1299.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, volume 34, 17864–17875. Curran Associates, Inc.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation. 16321–16330.

Feng, G.; Hu, Z.; Zhang, L.; and Lu, H. 2021. Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15501–15510. ISSN: 2575-7075.

He, J.; Deng, Z.; Zhou, L.; Wang, Y.; and Qiao, Y. 2019. Adaptive Pyramid Context Network for Semantic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7511–7520. ISSN: 2575-7075.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. ISSN: 2380-7504.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. ISSN: 1063-6919.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from Natural Language Expressions. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, 108–124. Cham: Springer International Publishing. ISBN 978-3-319-46448-0.

Huang, S.; Hui, T.; Liu, S.; Li, G.; Wei, Y.; Han, J.; Liu, L.; and Li, B. 2020. Referring Image Segmentation via Cross-Modal Progressive Comprehension. 10485–10494.

Hui, T.; Liu, S.; Huang, S.; Li, G.; Yu, S.; Zhang, F.; and Han, J. 2020. Linguistic Structure Guided Context Modeling for Referring Image Segmentation. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, 59–75. Cham: Springer International Publishing. ISBN 978-3-030-58607-2.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798. Doha, Qatar: Association for Computational Linguistics.

Kim, N.; Kim, D.; Lan, C.; Zeng, W.; and Kwak, S. 2022. ReSTR: Convolution-Free Referring Image Segmentation Using Transformers. 18145–18154.

Li, R.; Li, K.; Kuo, Y.-C.; Shu, M.; Qi, X.; Shen, X.; and Jia, J. 2018. Referring Image Segmentation via Recurrent Refinement Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5745–5753. ISSN: 2575-7075.

Liu, C.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; and Yuille, A. 2017. Recurrent Multimodal Interaction for Referring Image Segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1280–1289. ISSN: 2380-7504.

Liu, D.; Zhang, H.; Zha, Z.-J.; and Wu, F. 2019. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4672–4681. ISSN: 2380-7504.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11–20. ISSN: 1063-6919.

Margffoy-Tuay, E.; Pérez, J. C.; Botero, E.; and Arbeláez, P. 2018. Dynamic Multimodal Instance Segmentation Guided by Natural Language Queries. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, 656–672. Cham: Springer International Publishing. ISBN 978-3-030-01252-6.

Meinhardt, T.; Kirillov, A.; Leal-Taixé, L.; and Feichtenhofer, C. 2022. TrackFormer: Multi-Object Tracking With Transformers. 8844–8854.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. PMLR. ISSN: 2640-3498.

Shi, H.; Li, H.; Meng, F.; and Wu, Q. 2018. Key-Word-Aware Network for Referring Expression Image Segmentation. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, 38–54. Cham: Springer International Publishing. ISBN 978-3-030-01231-1.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, ; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2021. End-to-End Video Instance Segmentation with Transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8737–8746. ISSN: 2575-7075.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. CRIS: CLIP-Driven Referring Image Segmentation. 11686–11695.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language As Queries for Referring Video Object Segmentation. 4974–4984.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. S. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. 18155–18165.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10494–10503. ISSN: 2575-7075.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. 1307–1315.