LQMFormer: Language-aware Query Mask Transformer for Referring Image Segmentation

Nisarg A. Shah Vibashan VS Vishal M. Patel
Johns Hopkins University
snisarg812@gmail.com, {vvishnu2, vpatel36}@jhu.edu

Abstract

Referring Image Segmentation (RIS) aims to segment objects from an image based on a language description. Recent advancements have introduced transformer-based methods that leverage cross-modal dependencies, significantly enhancing performance in referring segmentation tasks. These methods are designed such that each query predicts different masks. However, RIS inherently requires a single-mask prediction, leading to a phenomenon known as Query Collapse, where all queries yield the same mask prediction. This reduces the generalization capability of the RIS model for complex or novel scenarios. To address this issue, we propose a Multi-modal Query Feature Fusion technique, characterized by two innovative designs: (1) Gaussian enhanced Multi-Modal Fusion, a novel visual grounding mechanism that enhances overall representation by extracting rich local visual information and global visual-linguistic relationships, and (2) A Dynamic Query Module that produces a diverse set of queries through a scoring network where the network selectively focuses on queries for objects referred to in the language description. Moreover, we show that including an auxiliary loss to increase the distance between mask representations of different queries further enhances performance and mitigates query collapse. Extensive experiments conducted on four benchmark datasets validate the effectiveness of our framework.

1. Introduction

Referring Image Segmentation (RIS) is a challenging multi-modal task aimed at segmenting specific objects in an image based on a textual description. This description often includes details about the object's actions, category, color, or position, as noted by [8, 21]. Compared to traditional semantic and instance segmentation, RIS requires a precise understanding of object locations and involves a comprehensive modeling of the visual-linguistic relationships within the global context. Additionally, RIS demands the extraction of high-quality visual features at a local level. The potential applications of RIS are extensive, particularly in language-driven human-computer interaction domains.

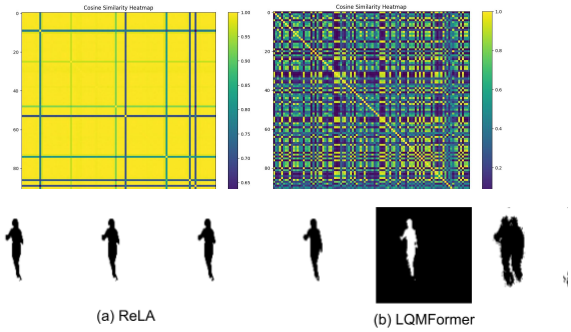

Figure 1. [Top] Heatmap visualization of cosine similarity between different query mask predictions of ReLA and LQMFormer , where the yellow region indicates a similarity near 1, and the dark blue region signifies a similarity of 0. [Bottom] Sample query mask prediction visualizations for ReLA and LQMFormer .

Traditional RIS methods have employed linear fusion approaches and Fully Convolutional Networks (FCNs) for feature learning in RIS tasks [21, 33]. However, the recent advancement of attention mechanisms has shifted the focus towards extracting richer visual-language representation, with recent techniques demonstrating the effectiveness of Transformers. These models excel at modelling long-distance dependencies, making them suitable for cross-modal fusion [45, 55]. Further, Vision Transformer (ViT) methods such as VLT [13], EFN [16], and LAVT [54], which are based on [14], have shown significant improvements in RIS performance. Their ability to capture the nuances of cross-modal dependencies is crucial, and through a series of attention mechanisms, these models efficiently process and integrate both visual and language inputs for precise object segmentation.

Despite their advantages, transformer-based approaches exhibit shortcomings in RIS mask prediction, with a notable problem being query collapse. This phenomenon occurs

when different queries within the transformer, intended to identify distinct attributes or portions of objects, converge on the same mask prediction. This is particularly problematic in referring image segmentation, which inherently requires a prediction of a single mask, where all queries are collectively trained to predict the same mask. We experimentally observed this phenomenon, as illustrated in Fig. 1. Here, we visualize the heatmap of the cosine similarity between mask predictions from ReLA [32] and our method. In ReLA, most mask predictions are almost identical, leading to significantly overlapping mask predictions, whereas our method demonstrates a diverse set of query embeddings and mask predictions. Hence, prior transformer-based methods train mask predictions for all queries to associate with a single, specific ground truth mask. This training approach can lead to query collapse, where the model is not penalized for producing the same prediction across different queries, thereby affecting its capability to identify diverse attributes of the image. Consequently, this limitation restricts the variety of mask predictions, diminishing the model's ability to generalize effectively in complex scenarios.

To address these challenges, we propose LQMFormer, which comprises two key components. Firstly, we introduce a Gaussian Enhanced Multi-Modal Fusion (GMMF) module, aimed at enhancing the visual grounding mechanism. This module is designed to extract fine-grained local visual details while simultaneously enhancing global visual-language relationships. The fusion of these modalities allows the model to construct a more comprehensive understanding of the scene, both globally and locally, thereby improving visual grounding. Secondly, in our Dynamic Query Module (DQM), we generate a diverse set of language-dependent queries through Language-Query Cross Attention (LQCA) and employ a scoring network to prioritize queries relevant to the referring expression. With an improved visual-grounded representation facilitated by GMMF and selective query enhancement via DQM, our model effectively addresses query collapse. Moreover, by incorporating an auxiliary loss that increases the representational distance between mask predictions, our method efficiently differentiates between queries, mitigating the issue of query collapse. Our extensive experiments across multiple benchmark datasets demonstrate the effectiveness of our framework. Our contributions can be summarized as follows:

- We introduce Gaussian enhanced Multi-Modal Fusion, which innovates the integration of local and global cross-modal information for more accurate visual grounding.
- We develop an effective Dynamic Query Module (DQM) that not only generates a diverse query set but also employs a scoring network to selectively focus on queries based on given language expressions with respect to the decoder's references.
- To mitigate query collapse, we propose an auxiliary regularization loss function that increases the representational distance between queries, significantly improving mask differentiation and preventing Query Collapse.
- We conduct comprehensive validation of the proposed framework with extensive testing on four benchmark datasets, demonstrating improvements in referring image segmentation performance.

2. Related Works

2.1. Referring Image Segmentation

Early works [29, 31, 45, 57] employed Convolutional Neural Networks (CNNs) [5, 21, 22, 31] and Recurrent Neural Networks (RNNs) [6, 31] to extract vision and language features. These features were then fused via sequential concatenation and convolution operations to predict the segmentation mask. In contrast, [58] proposed a two-stage network that initially uses Mask R-CNN[19] to predict categorical masks, followed by a selection of the relevant masks based on language descriptions. However, this approach exhibits limited capacity to capture the intricate relations between language and visual content. The advent of the Transformer [11, 14, 42, 46] in the vision community has led to the widespread adoption of Transformer-based architectures for extracting both visual and textual features [11, 24, 34]. VLT [13] uses cross-attention to produce query vectors from multi-modal features, which are then used to query images in the transformer decoder. Similarly, LAVT [54] shows early cross-modal fusion of features improves alignment. CRIS [49] adopts Transformer blocks to leverage the pre-trained CLIP model's [44] robust image-text alignment capabilities. Additionally, GRES [32] extends cross-attention in the Transformer decoder to explicitly model visual-language dependencies. However, these all transformer-based approaches require matching queries to corresponding mask instances. In the context of referring expression segmentation, where typically only a single mask is available, this can lead to all queries converging to a single mask, a phenomenon known as query collapse. Motivated by this, in our method, we try to produce a diverse set of query prediction by overall improving visual-grounding and conditioning with respect to language expression.

2.2. Vision and Language Representation Learning

Vision and Language Representation Learning focuses on understanding the semantics and alignment between vision and language for multimodal reasoning tasks [36, 47, 59]. This field has seen substantial progress in applications like visual question answering [2], Image captioning [48], Image-text retrieval [4], zero-shot classification [44], and Image segmentation [23]. Contrastive pre-training strategies [28, 44] using large-scale datasets effectively project multiple modalities into a single embedding space. In contrast, as discussed,

other methods [32, 54] create cross-modal interaction layers to fuse and comprehend multimodal features. The recent adoption of deep learning techniques in the frequency domain [9, 17, 39, 40] has gained attention for their global interaction capabilities. Specifically, [40] performs spectral-guided enhancement of the multi-scale vision-language features after feature extraction stage for Video Referring Segmentation. Drawing from these advancements, our approach integrates gaussian guidance within vision-language representation learning to facilitate enhanced local and global multi-modal interactions.

3. Methods

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a language expression $L = \{w_i\}_{i=1}^{N}$ with N words, our model predicts a pixel-wise mask M, which delineating the referred object.

3.1. Overview

An overview diagram of our approach is shown in Figure. 2. First, the language expression is encoded to extract high-dimensional language features $F_l \in \mathbb{R}^{N_t\times C}$, where C, N_t indicates the number of channels, and words in the language expression, respectively. The Gaussian-enhanced Multi-Modal Fusion (GMMF) module then extracts language-grounded visual features F_{vl}. Following [7], our approach incorporates a multi-scale strategy to exploit higher-resolution feature maps in the transformer decoder [7, 61]. We input language-aware queries Q_b' along with multi-scale outputs from the pixel decoder F_i, where $i \in \left\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\right\}$ of the original image. The Dynamic Query Module processes language features F_t to predict language-aware queries Q_b'. The updating of queries Q_b' sequentially utilizes resolutions $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$, applying a cycle of masked cross-attention (CA), self-attention (SA), and feed-forward network (FFN) operations L times within the decoder. Further, we map the final query outputs to an 'object' or 'no-object' two-dimensional space to enable 'no-object' predictions [32]. The final predicted mask is obtained at full resolution as the original image by decoding the pixel features $F_{\frac{1}{4}}$ with the einsum operation between Q and $F_{\frac{1}{4}}$, followed by upsampling operation.

3.2. Gaussian-enhanced Multi-Modal Fusion

After extracting language features F_l, we combine them with visual features using a four-stage hierarchical Swin Transformer [34] to obtain joint visual-language embedding. Each stage comprises Transformer layers (τ_i), multi-modal feature fusion modules (GMMF), and residual gating (ψ_i). At every stage, as Transformer layers (τ_i) enhance previous visual features (F_v), these features (F_v) combine with language features (F_l) via multi-modal modules (GMMF) to create multi-modal features (F_{vl}). Finally, we use gating units (ψ_{vl}) to weight and combine F_{vl} with F_v, yielding enhanced visual features with linguistic information. Residual gating also works well with pre-trained (only vision) transformer weights and allows simultaneous feature modulation with language without having to re-train from scratch.

3.2.1 Gaussian Enhanced Multi-Modal Fusion Module *in Visual Backbone*

The discrete Fourier transform (DFT) is extensively used to analyze the frequency representation of images. By transforming the signal into the frequency domain, characterized by global statistical properties, it plays a crucial role in various computer vision tasks [26, 51, 56]. Modulating signals on a point-wise basis in the Fourier domain alters the representation of inputs in the spatial domain [3, 40]. Furthermore, in the Fourier domain, low-frequency components usually correspond to the overall semantic content of the image, consistent with prior research [40, 52, 53, 56]. The Fourier transform is capable of disentangling global semantic information and structural coherence, with the semantic content predominantly found in the amplitude component [18, 27, 60]. These insights can be leveraged to create modules centered on Fourier-based methods, facilitating efficient and vital global interactions in multimodal understanding and improving visual grounding.

To inculcate the above observation, as illustrated in Fig. 3, following [40], the Gaussian enhanced Multi-Modal Fusion Module (GMMF) is proposed as a key component in our model. It employs an enhancement block after cross-attention operations to combine Gaussian-enhanced vision and language features. Given an input vision feature $F_{vi} \in \mathbb{R}^{N\times H\times W\times C_{vi}}$ from stage i and a language feature $F_l \in \mathbb{R}^{N\times T\times C_l}$, for cross-attention block, we first transform visual and language features into a common dimension C_i using separate 1×1 convolution layers. Subsequently, transform vision and language features are passed to an MHCA (Multi-Head Cross-Attention) module, where language features act as queries and vision features as keys and values. This operation results in the creation of a language-aware vision-shaped feature $F_{vl}^{'}$. We then transform the visual-language features using the Gaussian Enhancement Block, as shown in Fig. 3, formulated as follows:

$$F_{vl} = F_{\text{FFT}}^{-1}\Big(\text{Conv}\Big(FR\Big(\phi(F_{vl}^{'},\beta) * \mathcal{A}(F_{vl}^{'}) , \mathcal{P}(F_{vl}^{'})\Big)\Big)\Big) + F_{vl}^{'}$$

Here, $\mathcal{A}(F_{vl}^{'})$ and $\mathcal{P}(F_{vl}^{'})$ denote the amplitude and phase components of $F_{vl}^{'}$'s Fast Fourier Transform(FFT) [15, 43], $F_{\text{FFT}}(F_{vl}^{'})$. The $*$ represents low-pass filtering on $\mathcal{A}(F_{vl}^{'})$ using adaptive Gaussian filters $\phi(F_{vl}^{'},\beta)$, matching $F_{vl}^{'}$'s spatial dimensions and adjusting to input data via bandwidth β [40]. FR and $Conv$ signify feature reconstruction from amplitude and phase, and convolution operations, respectively.

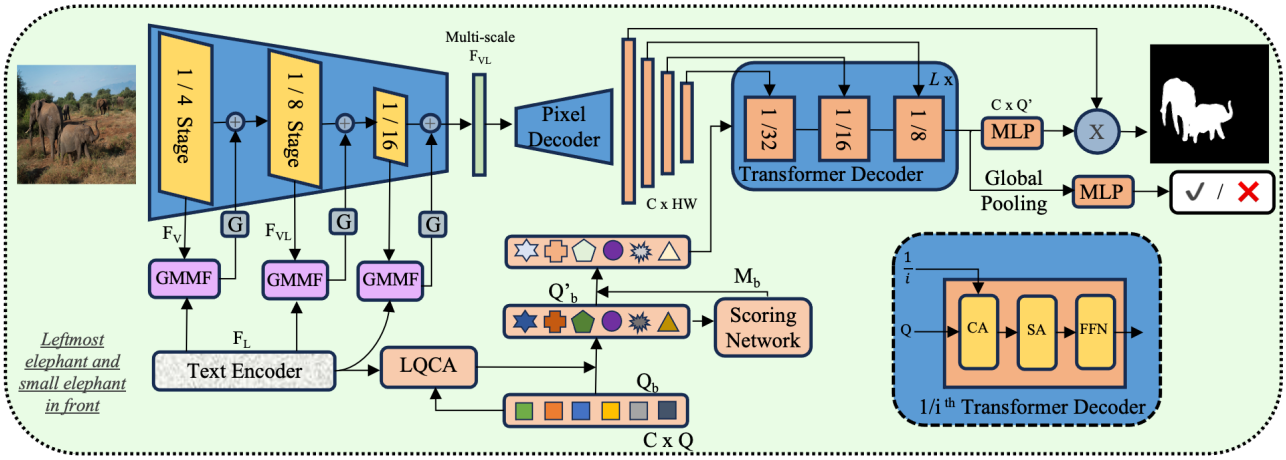

Figure 2. Overview of the LQMFormer model architecture. The model processes an image and a linguistic expression, extracting vision-language features F_{vl} through a Gaussian-enhanced Multi-modal Fusion module and language features F_L. The multi-scale vision-language features F_{vl} are then processed by into a pixel decoder. Concurrently, the Language Query Cross-Attention Module (LQCA) refines the Language-enhanced Query bank Q'_b with F_L. A Scoring Network assesses the relevance of each query in Q'_b by predicting soft-masks (M_b), which then modulate the queries fed into the transformer decoder. Finally, the model employs a global pooling layer and a Multi-Layer Perceptron (MLP) to predict the binary segmentation mask and determine the presence or absence of the referred object.

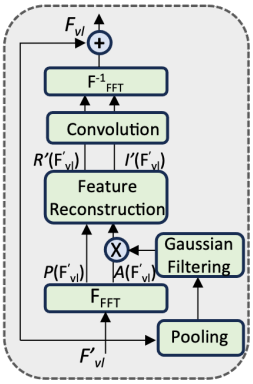

Figure 3. Architecture overview of the Gaussian Enhancement Module

Once F_{vl} is obtained, similar to [54], we combine the output from GMMF, F_{vl}, with that from the Transformer layers, F_v. This process employs a gating mechanism, ψ_{vl}, which learns a set of element-wise weight maps from F_{vl} to finely tune the scale of each element within F_{vl}. Subsequent to this adjustment, a residual combination of F_v and $F_{vl} \cdot \psi_{vl}$ is performed. The resultant output is then fed back into the vision backbone for further computation.

3.3. Dynamic Query Module

The Dynamic Query Module (DQM) takes the language feature F_l and a Query bank Q_b as inputs, the latter containing N_{qb} learnable queries. Figure 4(b) shows the initial step, where the attention between the language feature F_l and the N_{qb} query embeddings $Q_b \in \mathbb{R}^{N_{qb} \times C}$ is computed, resulting in N_{qb} attention maps:

$$A_{bi} = \text{softmax}(Q_b \sigma(F_l W_{ik})^T), \quad (1)$$

where W_{ik} is a $C \times C$ matrix of learnable parameters, and σ denotes the GeLU function [20]. The resulting $A_{bi} \in \mathbb{R}^{N_{qb} \times L}$ provides each query with a $1 \times L$ attention map, indicating its important linguistic relation in the language features. Following this, the language-supported query bank features are derived using these attention maps: $Q'_b = A_{bi} \sigma(F_l W_{iv})^T$, where W_{iv} is another $C \times C$ matrix of learnable parameters.

The variability in language expressions used to describe objects in images requires an approach that can adapt to different descriptions. This is because standard methods, like those seen in the initial transformer models (like in vanilla transformer [14]), struggle to capture the nuances in descriptions that include color, size, shape, and position. To address this, a larger set of adaptable queries is required. These queries are fine-tuned based on the input language, resulting in a refined set of queries, denoted as Q'_b. Each query in this refined set is then assessed with a scoring network, which outputs a soft-mask M_b. This soft-mask represents the relevance of each query. The final set of queries, Q_{final}, is obtained by combining M_b with Q'_b through a dot product, as shown in the equation below:

$$Q_{\text{final}} = M'_b \odot Q'_b \quad (2)$$

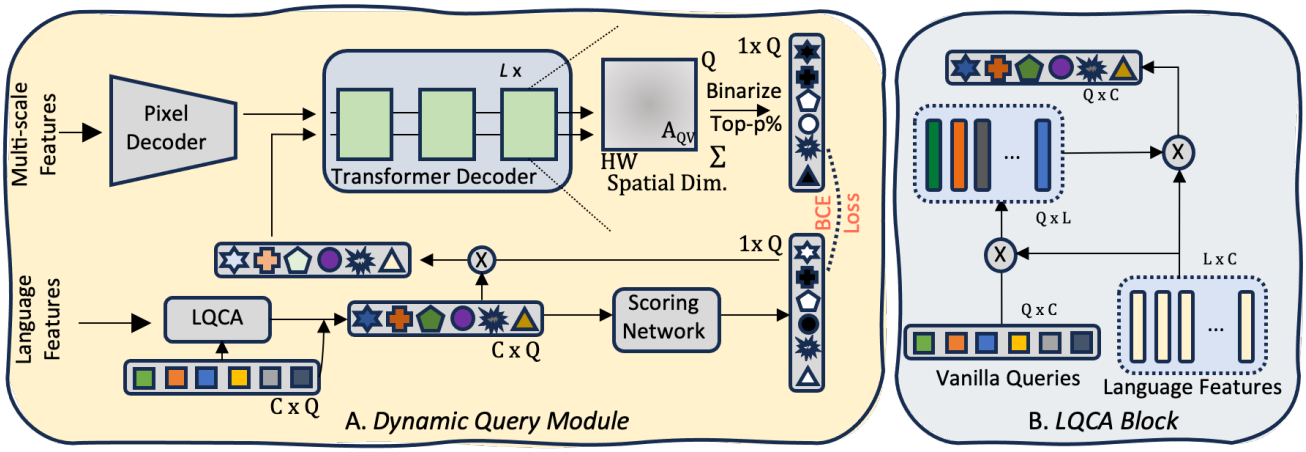

Figure 4. Architecture overview of the (A.) Dynamic Query Module and (B.) LQCA Block

3.3.1 Scoring Network

The scoring network in LQMFormer takes the Language-enhanced Query bank Q'_b as input and processes it through a sequence of linear layers. This network consists of four linear layers, with Layer Normalization [1] applied before the first layer. Each subsequent layer, except the last sigmoid, incorporates GELU [20] activation. The final output is a 1-dimensional logit $M_b \in \mathbb{R}^{N_{qb} \times 1}$, serving as a soft-mask to modulate Q'_b.

To train the scoring network for predicting a relevant soft-mask M_b effectively for the transformer decoder, we utilize its cross-attention map. This map is essential in identifying the subset of Queries most relevant to the referring expression and visual features, as highlighted during training. We sum up the attention maps from each decoder layer using Bilinear Interpolation (BI), thereby projecting attention $A_{qf} \in \mathbb{R}^{N_{qb} \times HW}$ directed towards these Queries. The process is mathematically described as follows:

$$A_{qf} = \sum_{l=1}^{L} \sum_{r=1}^{3} \text{BI}(\text{A}_{l,r}, HW) \tag{3}$$

Here, $\text{A}_{l,r} \in \mathbb{R}^{N_{qb} \times hw}$ represents the attention map from the l^{th} layer at the r^{th} resolution, with hw being the original dimension and HW the desired output dimension. After aligning each attention map to the common resolution HW, we then sum across this dimension to generate a likelihood map LM, as indicated in the subsequent equation:

$$LM = \sum_{i=1}^{HW} A_{qf}[:, i] \tag{4}$$

This likelihood map LM $\in \mathbb{R}^{N_{qb} \times 1}$, now captures the concentrated attention across all Queries, serves as a condition in refining the predictions of the scoring network. The scoring network is trained to predict the likelihood of each Query being included in the top-$\rho\%$ of the most referenced tokens. This process involves binarizing the likelihood map LM to retain only the top-$\rho\%$ Queries, creating a binarized version LM_{bin}. The training objective is then formulated as a Binary Cross-Entropy (BCE) loss between the predicted soft-mask M_b and the binarized likelihood map LM_{bin}. The BCE loss function is defined as:

$$\begin{aligned} \text{BCE Loss} = -\frac{1}{N_{qb}} \sum_{i=1}^{N_{qb}} [&LM_{\text{bin},i} \log(M_{b,i}) \\ &+ (1 - LM_{\text{bin},i}) \log(1 - M_{b,i})] \end{aligned} \tag{5}$$

Here, N_{qb} represents the number of Queries in Q'_b, and i indexes each Query. This loss function effectively trains the scoring network to align its predictions with the most salient Queries as determined by the decoder's cross-attention, thereby refining the model's accuracy in identifying and segmenting objects as described by the referring expressions.

3.4. Query-Mask Margin Loss

Query-Mask Margin Loss is designed to maintain a margin of separability between different query feature representations, an essential aspect for tasks with variable outputs. The loss function operates by initially computing pairwise differences and distances between mask embeddings within a batch. This computation yields a tensor of shape $B \times N \times N$, where B represents the batch size, and N is the number of queries. Subsequently, these distances for each embedding are sorted, with a particular focus on the second smallest value, which signifies the distance to the nearest neighbor. The margin loss is defined through the max function, ensuring a minimum distance of 1.0 between embeddings. The

mathematical formulation of QM-loss is as follows:

$$\text{QM-loss} = \max\left(0, 1.0 - \text{dist}\left(M_i, M_j\right)\right)$$

Here, dist (M_i, M_j) denotes the calculation of pairwise distances between the mask embeddings, represented as M.

4. Experiments and Discussion

4.1. Datasets and Implementation Details

Datasets: Our experiments are conducted on four primary benchmark datasets in Referring Image Segmentation: RefCOCO [57], RefCOCO+ [57], RefCOCOg [38], and GRES [32]. These datasets derive their images from MSCOCO [30]. The dataset details are as follows: RefCOCO (19,994 images, 50,000 objects, 142,209 expressions), RefCOCO+ (19,992 images, 49,856 objects, 141,564 expressions), RefCOCOg (26,711 images, 54,822 objects, 85,474 expressions), and GRES (19,994 images, 60,287 objects, 278,232 expressions). RefCOCO mainly focuses on expressions that specify object locations, while RefCOCO+ prioritizes object descriptions. RefCOCOg presents a higher challenge due to longer and more complex expressions (averaging 8.4 words compared to 3.5 in others). GRES broadens the scope of RefCOCO by including expressions that refer to multiple objects or no objects, thereby enhancing the problem's generalizability. We use both UMD split [41] and Google split [38] for RefCOCOg dataset.

Evaluation Metrics. For single-target object segmentation, we report our results using two kinds of metrics [13, 37, 49, 54]: overall IoU (oIoU) and Precision@X (P@X). oIoU divides the total intersection pixels by the total union pixels across all test images. Precision@X calculates the percentage of testing samples of which the model prediction has an IoU score higher than the threshold value X. To extend our evaluation to non-target and multi-target object segmentation, following [32], we include additional metrics: Sensitivity (N-acc.), Specificity (T-acc.), and Generalized IoU (gIoU). N-acc. evaluates the model's performance on identifying non-target samples, while T-acc. measures the impact of generalization on non-target samples on the performance of target samples. We opt for gIoU over oIoU due to the latter's bias towards larger objects. gIoU computes the mean IoU per image for all samples. In cases with no-target samples, IoU values are assigned as 1 for true positives and 0 for false negatives, ensuring an equitable metric for all object sizes.

Implementation Details. Our model uses Swin Transformer-B [34] as the visual encoder and BERT [11] as the language decoder. The transformer layer in Swin-B is initialized with classification weights pre-trained on ImageNet22K [10]. We use BERT implementation from the HuggingFace Transformer library [50]. It comprises 12 layers with a hidden size of 768, initialized using official pre-trained weights. We use our Transformer decoder proposed in Section 3.1 with L = 3 (i.e., 9 layers total) and 100 queries by default. The AdamW optimizer [35] is employed for optimization with an initial learning rate of 0.00001. Our model is trained over 100K iterations with a batch size of 48. Our image resizing is standardized to 480×480 pixels.

4.2. Comparison with the state of art methods

Referring Image Segmentation Table 1 presents the results for Referring Image Segmentation, and our proposed method, LQMFormer , produces competitive performance against existing state-of-the-art approaches across multiple standard benchmarks. On the RefCOCO dataset, our proposed method obtains the highest oIoU across all divisions (val, test A, and test B), surpassing recent leading methods such as VLT [13] and ReLA [32]. Notably, it shows a significant margin in the test B split, indicating enhanced capability in complex visual contexts. In the RefCOCO+ dataset, LQMFormer also demonstrates its effectiveness in the test A set. Specifically, in the test A split with a cIoU of 71.84%. In the Val and Test B splits, our proposed method achieves competitive performance against ReLA [32] and LAVT [54]. For the G-Ref dataset, which presents varied and longer referring expressions, our proposed method achieves the top performance in the $\text{val}_{(G)}$ split and maintains competitive results in the other splits. This performance is particularly notable over LAVT [54] in the $\text{val}_{(G)}$ split, highlighting its adaptability to diverse challenging datasets. While the margin of improvement offered by LQMFormer in classic RIS datasets may be comparatively smaller than that observed in GRES, the results signify that Gaussian-enhanced Multi-Modal Fusion and explicit modeling of Queries from Decoder are beneficial for the general Referring Image Segmentation setting.

Generalised Referring Image Segmentation In the GRES dataset, the LQMFormer model demonstrates notable improvements in Region Instance Segmentation (RIS), as shown in Table 2. The model achieves a generalized Intersection over Union (gIoU) score of 70.94% on the validation set, surpassing other models such as ReLA, LAVT, and CRIS by a margin of 7-12%. This performance indicates its effective segmentation across various object sizes. The object-level Intersection over Union (oIoU) score of 64.98% further underscores the model's segmentation capabilities, highlighting the effectiveness of the Improved Visual Grounding and Dynamic Query Selection module. Another important aspect of LQMFormer is its performance in no-target identification, with an accuracy score of 67.47%. This score is 11% higher than that of existing methods, marking significant progress in accurately identifying no-target samples—a noted challenge in RIS. The inclusion of the Dynamic Query Module (DQM) and the Top $p\%$ ratio significantly reduces the impact of the number of queries and mitigates query collapse. As a result, fewer but more effective queries make it easier to distinguish

Table 1. Results on classic RES in terms of oIoU. U: UMD split. G: Google split.

Methods	Visual Encoder	Textual Encoder	RefCOCO			RefCOCO+			G-Ref		
			val	test A	test B	val	test A	test B	$val_{(U)}$	$test_{(U)}$	$val_{(G)}$
MCN [37]	Darknet53	bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
VLT [12]	Darknet53	bi-GRU	67.52	70.47	65.24	56.30	60.98	50.08	54.96	57.73	52.02
ReSTR [25]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-	54.48
CRIS [49]	CLIP-R101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
LAVT [54]	Swin-B	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
VLT [13]	Swin-B	BERT	72.96	75.96	69.60	63.53	68.43	56.92	63.49	**66.22**	62.80
ReLA [32]	Swin-B	BERT	73.82	76.48	70.18	**66.04**	71.02	**57.65**	**65.00**	65.97	62.70
LQMFormer (ours)	Swin-B	BERT	**74.16**	**76.82**	**71.04**	65.91	**71.84**	57.59	64.73	66.04	**62.97**

Table 2. Comparison on GRES dataset.

Methods	val		No-target val	
	oIoU	gIoU	N-acc.	T-acc.
MattNet [58]	47.51	48.24	41.15	96.13
LTS [24]	52.30	52.70	-	-
VLT [12]	52.51	52.00	47.17	95.72
CRIS [49]	55.34	56.27	-	-
LAVT [54]	57.64	58.40	49.32	96.18
ReLA [32]	62.42	63.60	56.37	96.32
LQMFormer (ours)	**64.98**	**70.94**	**67.47**	**99.12**

Table 3. Ablation Study Results

Configuration	oIoU	gIoU
(a) Analysis on Dynamic Query module		
Ours	64.98	70.94
w/o score-based learning	62.75	67.16
w/o scoring module	62.02	65.48
w/o QM-loss	64.19	68.37
(b) Analysis on multi-modal fusion module		
Ours	64.98	70.94
w/o modality fusion module and QM-loss	49.68	52.34

Table 4. Ablation study of Query Numbers N_{qb} in Q_b. ‡: without scoring module. Top p% ratio is 20

N_q	oIoU	gIoU	Pr@0.7	Pr@0.8	Pr@0.9
20	57.22	58.31	69.15	60.26	22.47
50	62.48	64.53	72.02	61.79	31.58
100	64.98	70.94	75.03	65.51	34.05
100‡	62.02	65.48	73.15	62.87	31.55

between "target" and "no-target". The model's specificity (true accuracy) score of 99.12% also reflects its effectiveness in classifying target samples.

4.3. Ablation Study

In our comprehensive ablation studies, we evaluated the performance impact of various components of LQMFormer on the GRES dataset. These studies focused on the Dynamic Query Module, Query Numbers N_{qb} in Q_b, the Top p% ratio in Q_b, and the Grounding Module. The results helps understanding key insights into the model's functioning and optimization.

Table 5. Ablation study of Top p% ratio in Q_b.

$p\%$	oIoU	gIoU	N.acc.	Pr@0.7	Pr@0.8	Pr@0.9
20	64.98	70.94	67.47	75.03	65.51	34.05
50	63.85	67.76	63.25	74.29	64.42	33.58
100	61.33	63.77	58.92	72.45	61.88	31.21

Table 6. Ablation study of Grounding Module

Fusion	oIoU	gIoU	Pr@0.7	Pr@0.8	Pr@0.9
Baseline	49.68	52.34	55.27	41.38	14.72
Post-CA	60.22	61.89	74.15	64.44	32.67
S-Post-CA [40]	62.13	64.41	73.64	63.35	32.71
VL-Grounding	63.43	68.81	72.95	64.19	33.96
GMMF	64.98	70.94	75.03	65.51	34.05

Analysis on Dynamic Query Module: Table 3(a) presents the ablation results for the language query selection module. The removal of score-based learning significantly impacts scores as it limits the explicit training of Query scores from the scoring network. Also, notably, removing the scoring module and QM-loss leads to a decrease in oIoU score by 2.96% and 0.79%, respectively, highlighting its critical role in achieving optimal referring segmentation performance.

Query Bank number N_{qb} in Q_b: The ablation on Query Bank number demonstrated a notable impact on the model's performance as shown in Table 4. With an increase in N_q from 20 to 100, there was a consistent improvement in both oIoU from 57.22% to 64.98% and gIoU from 58.31% to 70.94%, as well as in Precision metrics. Particularly, $N_q = 100$ yielded the highest scores across all metrics, indicating an optimal balance in query numbers for effective segmentation. Thus, it optimal to have more number of queries with Top p% as 20.

Top p% Ratio in Q_b: The study on the Top p% ratio revealed that a 20% ratio achieved the best performance in

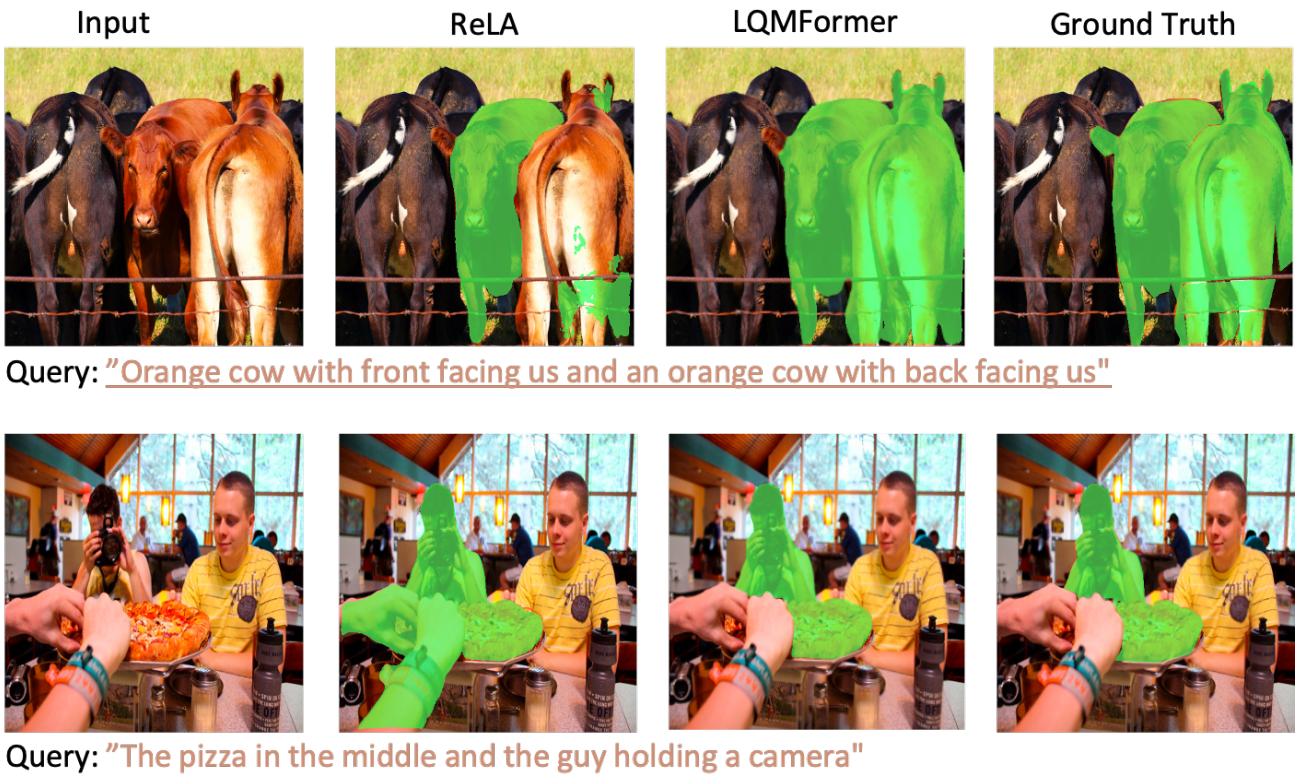

Figure 5. Qualitave Comparison of our model with LQMFormer with ReLA [32] on GRES dataset.

terms of oIoU, gIoU, and Precision metrics(Table 5). Higher ratios, like 50% and 100%, resulted in diminished performance as it resulted in more and more query collapse issues as many queries contribute towards the final mask generation. This finding indicates that a moderate selection of top-performing queries is crucial for balancing precision and recall.

Grounding Module: The ablation study of the Grounding Module, as shown in Table 3(b) and Table 6, highlights the effectiveness of Gaussian Enhancement for Referring Image Segmentation (RIS) tasks. The Gaussian enhanced Multi-Model Fusion (GMMF) method yields a improvement, specifically enhancing oIoU by 1.5%, from 63.43% in VL-Grounding to 64.98% in GMMF. This underlines the superior performance of Gaussian Filtering in the Fourier Domain over traditional cross-attention methods. Furthermore, compared to the Post-CA which registers an oIoU of 60.22%, the GMMF and VL-Grounding methods showcase substantial improvements in multi-modal representation within the backbone stage, with GMMF outperforming Post-CA by 4.76% and VL-Grounding by 3.21% in oIoU. The comparison with S-Post-CA [40], which aims at enhancing visual representation post-feature extraction, show the importance of integrating advanced grounding techniques early in the model's architecture for improved RIS performance. Overall, our proposed Gaussian enhanced Multi-Model Fusion (GMMF) improves the visual grounding compared to other methods, resulting in a improved referring segmentation.

5. Conclusion

In this paper, we address the problem of query collapse in Referring Image Segmentation by enhancing visual grounding and generating diverse query sets, and by modeling their relevance based on language expressions. Our proposed model, LQMFormer , incorporates Gaussian Enhanced Multi-Modal Fusion to improve visual grounding, and a Dynamic Query Module that not only generates a diverse set of queries but also employs a scoring network to selectively focus on queries based on given language expressions. Additionally, to mitigate query collapse, we introduce a novel loss function that enforces a margin of separation between query feature representations, facilitating improved representation without the need for additional supervision. The LQMFormer method notably surpasses state-of-the-art methods in both referring image segmentation and generalized referring image segmentation performance in most settings. For future work, we aim to extend the LQMFormer model's capabilities along with LLM and focus on effectively bridging LLM to better handle multi-modal contexts.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Ankan Bansal, Yuting Zhang, and Rama Chellappa. Visual question answering on image sets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 51–67. Springer, 2020. 2

[3] Glenn D Bergland. A guided tour of the fast fourier transform. *IEEE spectrum*, 6(7):41–52, 1969. 3

[4] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. 2

[5] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, pages 7454–7463, 2019. 2

[6] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*, 2019. 2

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[8] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. Imagespirit: Verbal guided image parsing. *ACM Transactions on Graphics (ToG)*, 34(1):1–11, 2014. 1

[9] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics, 2019. 2, 6

[12] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 7

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 6, 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 4

[15] Pierre Duhamel and Martin Vetterli. Fast fourier transforms: a tutorial review and a state of the art. *Signal processing*, 19 (4):259–299, 1990. 3

[16] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021. 1

[17] Lorenzo Giambagli, Lorenzo Buffoni, Timoteo Carletti, Walter Nocentini, and Duccio Fanelli. Machine learning in spectral domain. *Nature communications*, 12(1):1330, 2021. 3

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2

[20] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4, 5

[21] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124. Springer, 2016. 1, 2

[22] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, pages 4424–4433, 2020. 2

[23] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[24] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021. 2, 7

[25] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, pages 18145–18154, 2022. 7

[26] Sangrok Lee, Jongseong Bae, and Ha Young Kim. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2023. 3

[27] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023. 3

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[29] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, pages 5745–5753, 2018. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference,*

Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 6

[31] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, pages 1271–1280, 2017. 2

[32] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 2, 3, 6, 7, 8

[33] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *ICCV*, pages 4856–4864, 2017. 1

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2, 3, 6

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10437–10446, 2020. 2

[37] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 6, 7

[38] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 6

[39] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 3

[40] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 3, 7, 8

[41] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807. Springer, 2016. 6

[42] Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. *arXiv preprint arXiv:2403.12960*, 2024. 2

[43] Henri J Nussbaumer and Henri J Nussbaumer. *The fast Fourier transform*. Springer, 1982. 3

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[45] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, pages 38–54, 2018. 1, 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. pages 5998–6008, 2017. 2

[47] Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M Patel, and Ran Xu. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23539–23549, 2023. 2

[48] Yiyu Wang, Jungang Xu, and Yingfei Sun. End-to-end transformer based model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2585–2594, 2022. 2

[49] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 2, 6, 7

[50] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 6

[51] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021. 3

[52] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019. 3

[53] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, pages 264–274. Springer, 2019. 3

[54] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 1, 2, 3, 4, 6, 7

[55] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019. 1

[56] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[57] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 2, 6

[58] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 2, 7

[59] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 2

[60] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning*, pages 42589–42601. PMLR, 2023. 3

[61] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3