Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Language-Guided Hybrid Representation Learning for Visual Grounding on Remote Sensing Images

Biao Liu[1], **Xu Liu**[1*], **Lingling Li**[1], **Licheng Jiao**[1], **Fang Liu**[1], **Xinyu Sun**[1] and **Youlin Huang**[2]
[1]Xidian University
[2]East China Jiaotong University
{lb_email159, xuliu361}@163.com, llli@xidian.edu.cn, lchjiao@mail.xidian.edu.cn, {f63liu, sxy200102, 18270014345}@163.com

Abstract

Visual grounding (VG) refers to detecting the specific objects in images based on linguistic expressions, and it has profound significance in the advanced interpretation of natural images. In remote sensing image interpretation, visual grounding is limited by characteristics such as the complex scenes and diverse object sizes. To solve this problem, we propose a novel remote sensing visual grounding (RSVG) framework, named language-guided hybrid representation learning Transformer (LGFormer). Specifically, we designed a multi-modal dual-encoder Transformer structure called the adaptive multimodal feature fusion module. This structure innovatively integrates text and visual features as hybrid queries, enabling early-stage decoding queries to perceive the target position accurately. Then, the different modal information from the dual encoders is aggregated by hybrid queries to obtain the final object embedding for coordinate regression. Besides, a multi-scale cross-modal feature enhancement module (MSCM) is designed to enhance the self-representation of the extracted text and visual features and align them semantically. As for the hybrid queries, we use linguistic guidance to select visual features as the visual part and sentence-level features as the textual part. Finally, the LGFormer model we designed achieved the best results compared to existing models on the DIOR-RSVG and OPT-RSVG datasets.

1 Introduction

The task of visual grounding on remote sensing images (RSVG) is to locate specific objects according to the natural language expression [Sun *et al.*, 2022]. Unlike traditional object detection tasks [Girshick *et al.*, 2015; Carion *et al.*, 2020], visual grounding is not simply to classify and detect all the objects included in the training set, but to locate the unique objects described by natural language. Compared with natural images that have been discussed for a long time in the computer community, remote sensing (RS) images have more significant target scale changes and complex background interferences. Therefore, RSVG is extremely challenging. The area covered by an RS image can range from several square kilometers to thousands of square kilometers, and the scenes can include urban streets, port ships, mountains and rivers, etc. In such large-scale scenes, it is impossible for humans to quickly find some areas of interest manually, which also highlights the significance of this research. More importantly, RSVG has no professional requirements for users, allowing all users to easily implement target retrieval on RS images. This makes RSVG a broad research prospect in fields such as military reconnaissance, natural disaster surveillance, agricultural management, and urban development planning.

As an advanced research direction, RSVG is still underexplored. [Sun *et al.*, 2022] were the first to propose applying visual grounding to remote sensing images and introduced the RSVG dataset and the GeoVG method. GeoVG includes an image encoder, a language encoder, and a fusion module. This method uses the CNN network DarkNet-53 [Redmon and Farhadi, 2018], which incorporates an adaptive region attention module, as the image encoder, and the BERT [Devlin *et al.*, 2019] as the language encoder. The fusion module uses an attention network to integrate textual information into visual features, thereby obtaining fused visual features that can be used for target coordinate regression. [Zhan *et al.*, 2023] proposed a new method for RSVG called MGVLF and a new dataset called DIOR-RSVG, constructed based on the DIOR [Li *et al.*, 2020] dataset. MGVLF takes into account the significant size differences of target objects in remote sensing images. A single-scale feature map struggles to effectively capture the detailed information of targets of different sizes simultaneously. Therefore, it uses multi-scale visual feature maps to provide visual information at different levels. MGVLF also extracts multi-granularity text features, including sentence-level features that provide global contextual information and word-level features that provide local semantic information. In the recent work LQVG [Lan *et al.*, 2024], the authors argue that the target object occupies a small spatial area in remote sensing images, and its visual representation is limited in the cross-modal features obtained by concatenating visual and textual features. This makes it difficult for a single learnable token to effectively gather target information through the self-attention mechanism in the cross-modal fusion module. Thus, they proposed a language

* Corresponding author.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

query-based Transformer framework, using multiple repeated sentence-level text embeddings as queries to aggregate target object information from multi-scale visual features, and using this as the final object embedding, thereby improving localization accuracy. The work of LPVA [Li *et al.*, 2024] believes that past methods relied solely on visual information when extracting features through the visual backbone, without considering the potential correlation between visual and textual information. This may cause attention drift during the feature extraction process, leading to the extraction of visual features of objects that are inconsistent with the textual description. To address this, they designed a progressive attention module and a multi-level feature enhancement decoder, which dynamically adjusts visual features using linguistic information, successfully solving the problem of attention drift.

Despite the significant progress made by the aforementioned methods in the RSVG task, the performance in practical applications still needs improvement. Moreover, we believe that existing methods are not efficient in utilizing the extracted features when obtaining object embeddings, resulting in poor performance of object embeddings during regression. Simply using a learnable representation to initialize object embeddings is insufficient to fully learn the complete target localization information. Recently, some methods have used the embedding of a single modality as a query to fuse information from another modality's features, and then used this as the object embedding. However, this still struggles to provide sufficient fusion prior information for the initial query, making it difficult to handle the differences and interferences between different modality features in complex remote sensing image scenarios, resulting in weak robustness.

In light of this, the motivation of this study is to propose a novel method that improves the quality and robustness of object embeddings by more efficiently integrating multimodal features, thereby enhancing the performance of RSVG. In this paper, we propose a language-guided hybrid representation learning Transformer framework (LGFormer) using hybrid queries for RSVG. Specifically, our model includes an image encoder, a text encoder, a multi-scale cross-modal feature enhancement module (MSCM), a language-guided visual feature filtering module, and a adaptive multimodal feature fusion module. The image encoder uses a convolutional neural network (CNN) and a vision transformer (ViT) network in parallel to extract multi-scale features from the image. BERT extracts multi-granularity features from text, including sentence-level features and word-level features. After passing through the MSCM module, the extracted text and visual features will be refined into embedding vectors and aligned in the semantic space. Next, we will use the text embeddings to filter the visual embeddings, and the filtered visual embeddings will be combined with sentence-level embeddings to serve as hybrid queries for the adaptive multimodal feature fusion module. Our adaptive multimodal feature fusion module is similar to the traditional detect transformer structure, but it includes an additional encoder. In this way, hybrid queries can simultaneously interact and fuse with text and visual bimodal features during the decoding stage, improving the efficiency of utilizing key features from different modalities. Due to the prior characteristics of the content initialized by the hybrid queries, it inherently possesses the ability to perceive the target. Ultimately, the object embeddings we obtained achieved higher coordinate regression accuracy.

Overall, the main contributions of this paper consist of the following three points.

1. The language-guided hybrid representation learning Transformer (LGFormer) is proposed for the task of remote sensing visual grounding. It allows decoding query to improve the efficiency of utilizing the extracted features from different modalities, thereby obtaining more accurate object embeddings.
2. The hybrid query is designed to initialize the decoding query. The hybrid query consists of sentence-level features and multi-scale visual features, providing rich prior information to ensure its semantic understanding ability and cross-domain adaptability during the decoding process.
3. A language-guided visual feature filtering method is proposed. By using textual features to select visual features, we ensure that the multimodal information in the hybrid query is semantically highly relevant, thereby correctly guiding the aggregation of contextual information of the referred objects.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 elaborates on the proposed LGFormer method. Section 4 provides a comprehensive overview of the experimental setup and results. Finally, we conclude the paper in Section 5.

2 Related Work

2.1 Visual Grounding on Natural Image

The following will review visual grounding on natural images from three aspects: two-stage methods, one-stage methods, and Transformer-based methods.

1) Two-Stage Method: The two-stage method typically divides visual grounding into two independent processes: region generation and text-region matching. This is somewhat similar to traditional two-stage object detection methods [Girshick *et al.*, 2015; Girshick, 2015; Ren *et al.*, 2017]. The first stage of region generation is generally achieved using pre-trained object detectors [Ren *et al.*, 2017; Redmon, 2016; Liu *et al.*, 2016]. The related research in the second stage is the focus of this direction, where language expressions are used to select the best-matched region among the many generated. Earlier methods [Nagaraja *et al.*, 2016; Wang *et al.*, 2016] generally achieved good matching results by optimizing feature embedding networks with maximum margin ranking loss to maximize the similarity between object-query pairs. SCRC [Hu *et al.*, 2016] takes text queries, candidate regions, their spatial configurations, and global context as input, generating scores for each candidate region. MattNet [Yu *et al.*, 2018] introduces a modular design and improves the accuracy of object localization by better modeling language descriptions related to subjects, locations, and relationships. Recent research has further improved the two-stage method by better modeling object relationships [Yang *et al.*, 2019a;

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

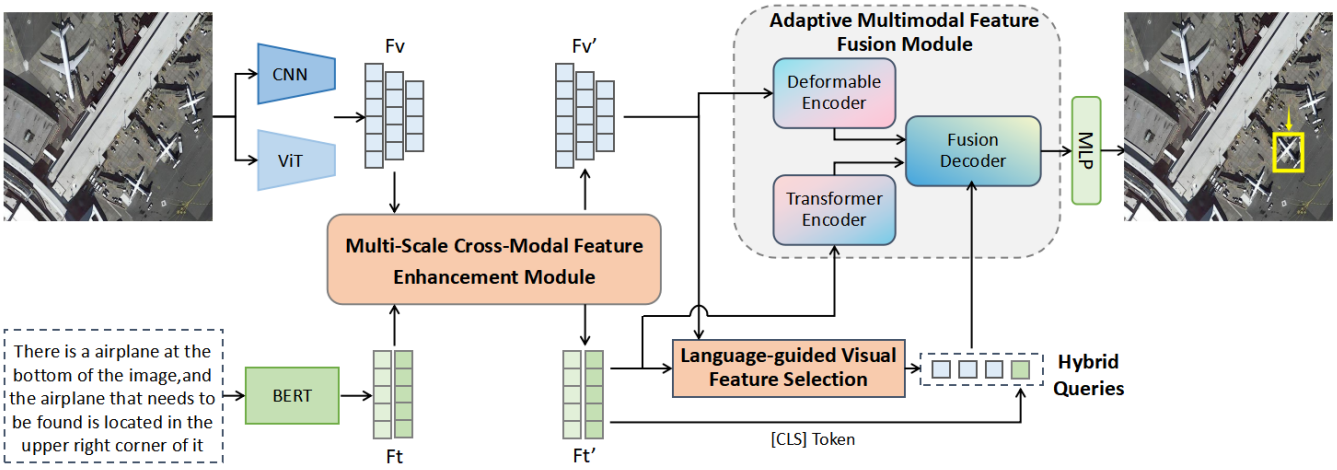

Figure 1: Overview of the proposed LGFormer framework. The MSCM module aligns the extracted multi-scale visual features and multi-granularity text features, then uses the text features as a guide to filter the visual features. The filtered results are then combined with sentence-level features to form hybrid queries. In the adaptive multimodal feature fusion module, hybrid queries are used for information retrieval and aggregation of multimodal features to obtain the final object embeddings for prediction.

Yang *et al.*, 2020a; Wang *et al.*, 2019] or utilizing phrase co-occurrence [Bajaj *et al.*, 2019; Dogan *et al.*, 2019]. However, the drawbacks of the two-stage method are also evident. Its filtering operation on a large number of invalid candidate regions leads to a significant waste of computational resources and time. Moreover, the performance of the second stage is directly constrained by the quality of the candidate regions provided by the first stage.

2) One-stage methods: One-stage methods support directly extracting information and predicting the coordinates of targets on visual features after densely integrating text features. The early FAOA [Yang *et al.*, 2019b] integrated text embeddings into the YOLOv3 detector [Redmon and Farhadi, 2018], achieving end-to-end visual grounding. RCCF [Liao *et al.*, 2020] uses text features as filters to verify and filter visual features. Recently, ReSC [Yang *et al.*, 2020b] proposed a recursive subquery construction framework, which improves one-stage visual grounding by addressing the localization limitations of long and complex queries.

3) Transformer-based methods: Transformer [Ashish, 2017] has achieved significant success in both natural language processing and computer vision. TransVG [Deng *et al.*, 2021] first proposed a transformer-based visual grounding method, achieving cross-modal fusion through a transformer encoder. It concatenates query tokens with visual and textual features, which are then fed into the fusion module to aggregate useful information for coordinate regression. TransVG++ [Deng *et al.*, 2023] further transformed the model into a pure transformer structure, unifying visual feature encoding and multimodal fusion, thereby improving the performance. In the VLTVG [Yang *et al.*, 2022], the authors developed a visual-language verification module before the object localization stage to adjust the relationship between visual features and text features. QRNet [Ye *et al.*, 2022] used a query-aware dynamic attention mechanism and multi-scale fusion to adjust the intermediate features in the visual backbone, thereby addressing the issue of inconsistency between the visual features extracted from the visual backbone and the features truly required for multimodal reasoning.

2.2 Visual Grounding on RS Images

The research of RSVG started late and is more challenging. Due to complex size variations and background interference, it is more difficult to separate the target's contours from the image. Sun et al. [Sun *et al.*, 2022] first proposed the RSVG task, and introduced the GeoVG method and the RSVG dataset. GeoVG enhances its spatial location understanding capability through an adaptive region attention module. [Zhan *et al.*, 2023] proposed the MGVLF and the DIOR-RSVG dataset. MGVLF integrates multi-scale visual features with multi-granularity textual features to help improve localization performance. Recently, [Lan *et al.*, 2024] proposed a multimodal Transformer using language queries, where the language queries use multiple sentence-level feature embeddings instead of a single learnable token to aggregate the contextual information of the referenced objects. [Li *et al.*, 2024] argue that existing methods lack interaction with textual information when extracting visual features, leading to attention drift. Therefore, they incorporated language guidance into the visual backbone of their LPVA, enhancing precise attention to the guided objects, and achieved good results on their proposed OPT-RSVG dataset.

3 Method

3.1 Overview

As shown in Figure 1, the proposed LGFormer consists of four main parts: visual and text encoders, a multi-scale cross-modal feature enhancement module, a language-guided visual feature filtering module and a adaptive multimodal feature fusion module. During the inference process, the encoder extracts features from the given image-text pairs, obtaining

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

the corresponding multi-scale visual features F_v and multi-granularity text features F_t. Subsequently, the extracted features will be aligned in the semantic space through the cross-attention mechanism in the MSCM module, resulting in further refined visual and textual features. Next, we use the refined textual features as language guidance to filter the refined visual features in the language-guided visual feature filtering module, and concatenate the output with the sentence-level features to form hybrid queries. Meanwhile, the refined visual and text features are respectively fed into the two different modality encoder branches of the adaptive multimodal feature fusion module. Their outputs, along with the hybrid queries, are decoded in the fusion decoder to obtain the object embeddings of the referenced objects. Finally, the object embeddings are used for category prediction and bounding box coordinate regression through the MLP head.

3.2 Visual-Text Fundamental Feature Representation

The feature extraction encoder consists of two parts: the visual encoder and the text encoder.

1) Visual Encoder: We adopted a parallel combination of CNN model with ResNet [He *et al.*, 2016] and vision transformer (ViT) [Dosovitskiy *et al.*, 2021] as the main backbone for visual extraction. ViT encodes image patches like it processes sequence data, thereby better capturing global information and long-range dependencies within the image. This is particularly effective for understanding large-scale features such as land cover distribution and topography in remote sensing images. CNN, due to its convolutional structure with local receptive fields, is better at extracting local features. Specifically, we obtained the outputs of the last four layers of ViT and the last three layers of ResNet. Additionally, we performed another downsampling with a stride of 2 on the output of the last layer of ResNet to obtain the output of the fourth layer.

2) Text Encoder: For language expressions, we use a pre-trained BERT with 12 hidden layers to obtain multi-granularity text information. Specifically, we obtain the output of the last layer of BERT as the sentence-level features F_t^s, and the average of the outputs of the last four layers as the word-level features F_t^w. Sentence-level features are responsible for providing overall contextual information, helping to understand the approximate location of the target object within the scene. Word-level features focus on specific object names, colors, shapes, and other attribute words in the text, used for precise matching of targets in RS images. F_t represents the final extracted text embeddings: $F_t = [F_t^s, F_t^w]$.

3.3 Multi-Scale Cross-Modal Feature Enhancement

As shown in Figure 2, in the multi-scale cross-modal feature enhancement module (MSCM), we simply use two cross-attention modules to perform feature fusion from vision to text and from text to vision, respectively. Since the principles of the two cross-modal fusions are the same, we will explain the cross-fusion from vision to text as an example. First, the multi-scale visual embeddings are projected into Query (Q) through the W^Q linear projection, while the multi-granularity

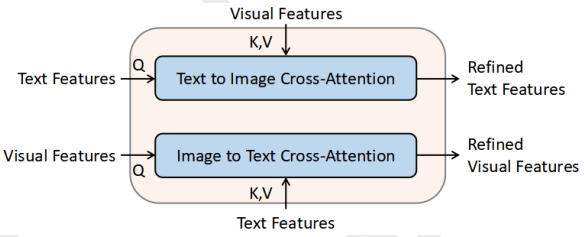

Figure 2: Illustration of the proposed multi-scale cross-modal feature enhancement (MSCM) module. The two cross-attention layers contained in the MSCM module are both implemented using a multi-head attention mechanism.

text embeddings are projected into Key (K) and Value (V) through the W^K and W^V linear projections, respectively. For the i-th iteration, the attention mechanism calculates the cross-modal weight A_{vt}^i using the following formula:

$$A_{vt}^i = \text{Softmax}\left(\frac{Q \cdot (K)^T}{\sqrt{d_k}}\right), \tag{1}$$

subsequently, we use the computed ${A_{vt}}^i$ and multiply them with Value to obtain the attention output. Finally, we update the visual embedding F_v^i through multiplication, thereby obtaining the refined visual feature $F_v^{i'}$.

$$F_v^{i'} = A_{vt}^i V \cdot F_v^i. \tag{2}$$

3.4 Language-Guided Visual Feature Filtering

To construct the hybrid query for decoding, we use sentence-level embeddings and multi-scale visual embeddings to initialize the query. However, due to the large number of embedding vectors obtained after refining visual features, it is impractical to use all visual embeddings as the visual component of the subsequent hybrid query. Therefore, we use text features to filter them. Specifically, we calculate the similarity between text embeddings and all visual embeddings using the dot product, and then find the top k visual embeddings in descending order of similarity to form the visual part. The calculation formula is as follows:

$$I_k = Top_k(Max^{(-1)}(\mathbf{X}_v\mathbf{X}_t^\top)), \tag{3}$$

here, I_k represents the indices of the top k visual embeddings. Top_k refers to the operation of obtaining the indices of these embeddings in the visual embedding tensor based on the input content. $Max^{(-1)}$ and the symbol ^{T} represent the operations of taking the maximum along the -1 dimension and transposing the matrix, respectively. The input X_v of this calculation formula represents visual features, and the input X_t represents text features.

3.5 Adaptive Multimodal Feature Fusion

Adaptive multimodal feature fusion module aims to efficiently obtain object embeddings to assist with coordinate regression. To enrich the prior information of the initial queries, we construct hybrid queries by combining sentence-level embeddings and k filtered visual embeddings. Note that we originally used the hidden representations from the last layer out-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

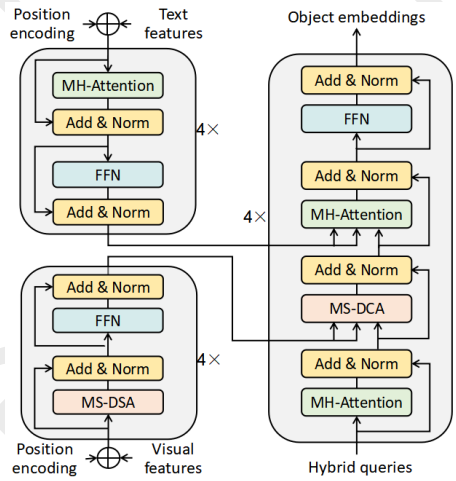

Figure 3: Detailed Structure of Adaptive Multimodal Feature Fusion Module. The adaptive multimodal feature fusion module consists of a text feature encoder, a visual feature encoder, and a cross-modal decoder.

put of BERT as sentence-level embeddings, but when forming the hybrid queries, we only select the [CLS] embedding from this output as the text part of the queries.

The detailed structure is shown in Figure 3. The two encoders receive refined visual and textual features as input and further optimize these multimodal features, allowing the model to focus on key parts. Notably, considering that the visual encoder branch and the subsequent decoder input visual features are all multi-scale, we introduce the multi-scale deformable attention mechanism (MS-DA) from Deformable-DETR [Zhu *et al.*, 2020]. MS-DA combines the sparse spatial sampling of deformable convolution [Dai *et al.*, 2017] and multi-scale feature fusion, greatly meeting the demand for multi-scale information in RSVG. Additionally, both the encoder branches and the decoder are composed of four stacked blocks each. Specifically, each block of the visual encoder branch consists of an MS-DSA module and an FFN module. Similarly, by simply replacing the MS-DSA module with a standard multi-head self-attention module, we can obtain the block for the text encoder branch. The decoder block consists of a multi-head self-attention module, two cross-attention modules, and an FFN module. The two cross-attention modules are used for cross-modal interaction between hybrid queries and text embeddings, as well as visual embeddings. Guided by prior knowledge, the hybrid queries interact with features of different modalities in the decoder, aggregating key information of referred objects. Ultimately, we obtain efficient object embeddings for target localization.

3.6 Loss Function

The model ultimately predicts object categories and regresses box coordinates based on object embeddings. Therefore, our loss function includes both classification loss and regression loss.

The focal loss function is employed in our classification loss calculation. It utilizes a dynamic scaling factor to progressively lessen the importance of samples that are simple to differentiate during the training process. This allows the model to rapidly concentrate on the more challenging samples that are difficult to distinguish. The mathematical expression for focal loss is given below:

$$L_{focal}(p_i) = -\alpha_i(1-p_i)^{\gamma}\log(p_i) \tag{4}$$

In the above formula, p_i represents the model's predicted probability for each category, α_i is the balance factor used to balance the weights of positive and negative samples, and γ is the dynamic adjustment factor.

For the calculation of regression loss, we use a combination of L1 loss and GIoU loss [Rezatofighi *et al.*, 2019]. L1 loss can directly measure the distance difference between the predicted box and the ground truth box, allowing for rapid convergence in the early stages of training. GIoU loss is a loss function based on the overlap of bounding boxes, which can better reflect the geometric relationship between the predicted box and the ground truth box. Combining the two can simultaneously leverage the rapid convergence characteristics of L1 loss and the geometric constraint capabilities of GIoU loss, allowing the model to quickly optimize in the early stages of training and better adjust the shape and position of the bounding boxes in subsequent stages. Specifically, the calculation formulas for L1 and GIoU losses are as follows:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{5}$$

$$L_{GIoU} = 1 - GIoU, \tag{6}$$

$$GIoU = IoU - \frac{C-(A\cup B)}{C}, \tag{7}$$

here, y_i represents the coordinates of the ground truth box, and $\hat{y}_i$ represents the coordinates of the predicted box. In the GIoU calculation formula, A and B represent the areas of the ground truth box and the predicted box, respectively, and C represents the area of the smallest enclosing rectangle of the ground truth box and the predicted box. IoU is the intersection over union of the ground truth box and the predicted box.

4 Experiments

4.1 Datasets

1) DIOR-RSVG: The DIOR-RSVG dataset [Zhan *et al.*, 2023] was constructed based on DIOR [Li *et al.*, 2020]. This dataset consists of 17,402 RS images of size 800×800 and 38,320 corresponding language expressions. There are a total of 20 object categories, and the average number of words in the language expressions is 7.47. In the released version of DIOR-RSVG we used, the split ratios for the training set, validation set, and test set image-text pairs are 70%, 10%, and 20%, respectively.

2) OPT-RSVG: OPT-RSVG [Li *et al.*, 2024] includes a wider range of target scales, not only with a similar number of small objects as DIOR-RSVG, but also with a higher proportion of small and large targets, increasing the challenge of remote sensing object recognition. The OPT-RSVG dataset

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Methods	Venue	Visual Enc.	Text Enc.	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
TransVG	CVPR'21	ResNet-50	BERT	72.41	67.38	60.05	49.10	27.84	63.56	76.27
LBYL-Net	CVPR'21	DarkNet-53	BERT	73.78	69.22	65.56	47.89	15.69	65.92	76.37
QRNet	CVPR'22	Swin-S	BERT	75.84	70.82	62.27	49.63	25.69	66.80	83.02
VLTVG	CVPR'22	ResNet50	BERT	69.41	65.16	58.44	46.56	24.37	59.96	71.97
MGVLF	TGRS'23	ResNet50	BERT	75.98	72.06	65.23	54.89	35.65	67.48	78.63
LQVG	TGRS'24	ResNet50	BERT	83.41	81.03	75.91	65.52	43.53	74.02	82.22
LPVA	TGRS'24	ResNet50	BERT	82.27	77.44	72.25	60.98	39.55	72.35	85.11
Ours	-	ResNet50(&)ViT	BERT	**88.81**	**86.73**	**81.76**	**70.16**	**46.21**	**78.72**	**85.65**

Table 1: Compared with state-of-the-art methods on the DIOR-RSVG test set

Methods	Venue	Visual Enc.	Text Enc.	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
TransVG	CVPR'21	ResNet-50	BERT	69.96	64.17	54.68	38.01	12.75	59.80	69.31
LBYL-Net	CVPR'21	DarkNet-53	BERT	70.22	65.39	58.65	37.54	9.46	60.57	70.28
QRNet	CVPR'22	Swin-S	BERT	72.03	65.94	56.90	40.70	13.35	60.82	75.39
VLTVG	CVPR'22	ResNet50	BERT	71.84	66.54	57.79	41.63	14.62	60.78	70.69
MGVLF	TGRS'23	ResNet50	BERT	72.19	66.86	58.02	42.51	15.30	61.51	71.80
LQVG	TGRS'24	ResNet50	BERT	86.09	83.17	76.88	62.01	29.80	73.94	78.16
LPVA	TGRS'24	ResNet50	BERT	78.03	73.32	62.22	49.60	25.61	66.20	76.30
Ours	-	ResNet50(&)ViT	BERT	**86.81**	**84.57**	**79.07**	**65.01**	**33.97**	**75.35**	**79.94**

Table 2: Compared with state-of-the-art methods on the OPT-RSVG test set

contains 25,452 RS images and 48,952 pairs of language expressions, with the language expressions divided into English and Chinese versions. Additionally, this dataset includes 14 object categories. The official division ratios for the training set, validation set, and test set provided by the OPT-RSVG dataset are 40%, 10%, and 50%, respectively. Note that we only used the English version of the language expressions in our experiments.

4.2 Implementation Details

The LGFormer we proposed is implemented using Pytorch, just like the other deep learning models we compared it with. In the model's image encoder, we use the pre-trained ResNet50 as the CNN branch and stack 8 transformer encoder layers to form the ViT branch. At the same time, we use the pre-trained BERT-base as the text encoder. The hidden dimensions of the feature extraction backbones for both vision and text are 768. As for the adaptive multimodal feature fusion module, our encoder and decoder are each composed of four stacked layers, with a hidden dimension of C=256. The number of hybrid queries N=10 (one sentence-level embedding and k=9 visual embeddings).

During the training process, we conducted distributed training on four RTX 3090 GPUs (24 GB VRAM). On the DIOR-RSVG dataset, we conducted a total of 40 training epochs. We froze the text encoder and the CNN branch in the visual encoder during the first 20 training epochs, and only froze the text encoder during the last 20 training epochs. For the OPT-RSVG dataset, we fine-tuned the model trained on the DIOR-RSVG training set for 15 epochs. In all training sessions, we set the batch size to 2 per GPU and used AdamW as the optimizer. The initial learning rate for the text encoder BERT is 1e-5, and the rest are 1e-4. It is worth noting that during the training process of the two datasets, the size of each image is randomly resized to the range of [480, 560, 640, 720, 800], but this operation is not performed during inference.

4.3 Evaluation Metrics

To evaluate the effectiveness of our model, we chose the same evaluation method as most RSVG papers. When the threshold of the intersection area to the union area between the predicted box and the ground-truth box surpasses a specific threshold, the predicted box is deemed accurate. The IoU thresholds we used range from 0.5 to 0.9, denoted as Pr@0.5, Pr@0.6, Pr@0.7, Pr@0.8, and Pr@0.9. In addition, we also use meanIoU and cumIoU as evaluation metrics, and their specific calculation formulas are as follows.

$$meanIoU = \frac{1}{N}\sum_{t}\frac{I_t}{U_t}, \tag{8}$$

$$cumIoU = \frac{\sum_t I_t}{\sum_t U_t}, \tag{9}$$

here, t represents the index of the image-text data and N is the size of the data. I_t and U_t represent the intersection and union of the areas of the predicted box and ground-truth box, respectively.

4.4 Compared with State-of-the-Art Method

We compared the performance of the proposed method with other SOTA methods on DIOR-RSVG and OPT-RSVG. The best and second-best performances are highlighted in bold

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

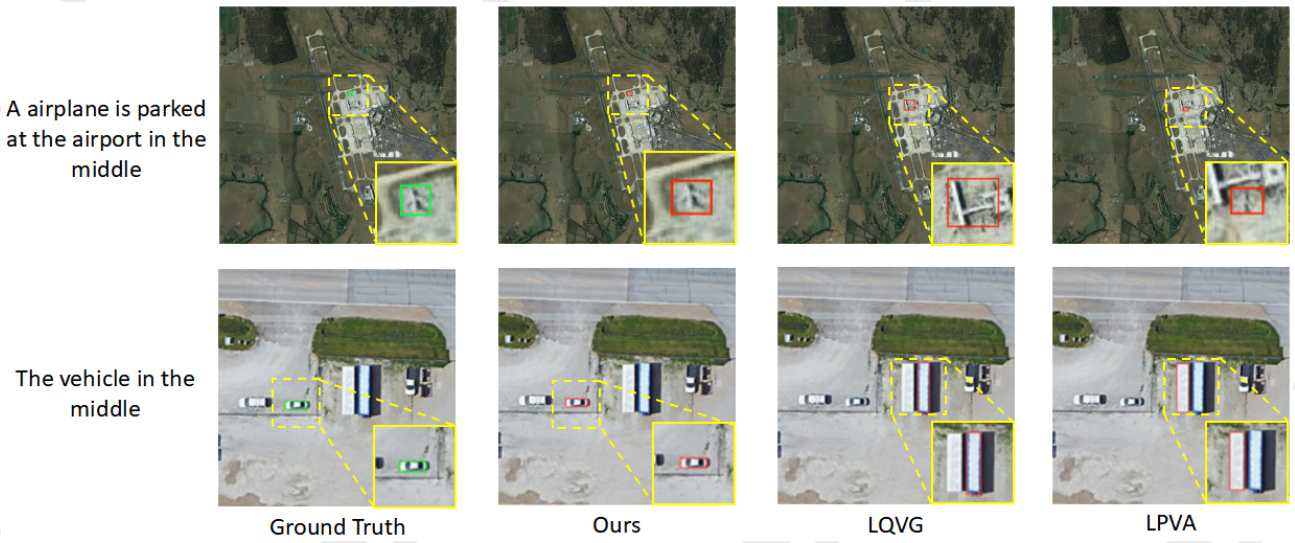

Figure 4: Visualization results comparison between our LGFormer and other existing SOTAs on DIOR-RSVG. GT is indicated by green boxes, while the output boxes of our LGFormer, LQVG, and LPVA are indicated by red boxes.

and underlined, respectively. Note that the performance data of other deep learning models are cited from reference [Li *et al.*, 2024], which is also the latest SOTA method we compared against. To ensure fairness, the hardware platforms used for training and testing were consistent with those in reference [Li *et al.*, 2024].

As shown in Table 1 and Table 2, our LGFormer achieved the best performance on both the datasets, validating the effectiveness of our method. Compared to latest works LQVG and LPVA, the proposed method achieves highest accuracy on all evaluation metrics across the DIOR-RSVG and OPT-RSVG datasets. In early transformer-based methods such as TransVG, QRNet, VLTVG, and MGVLF, a single learnable query is used to aggregate key contextual information of objects from visual and textual features in cross-attention. Recent works LQVG and LPVA use language modality-based features to initialize queries. In our work, we proposed a adaptive multimodal feature fusion module using hybrid queries, combining sentence-level features and filtered visual features as initial queries, and demonstrated the superior performance of this framework through experimental results.

Figure 4 shows the visualization of the comparison results between LGFormer and other SOTA methods. From the first set of comparison images, it can be seen that only our LGFormer successfully located the correct target. LQVG incorrectly identified the platform as the airplane, while the airplane located by LPVA was in a different position than required by the language description. From the second set of comparison images, it can be seen that, apart from our model correctly locating the target, both LQVG and LPVA incorrectly identified the middle container as the target vehicle. In summary, it is clear that our LGFormer performs the best. This is because the hybrid queries we used contain rich prior information, which is beneficial in guiding the queries to retrieve and aggregate the visual and text features during the subsequent decoding process. Moreover, the proposed adaptive multimodal feature fusion module further optimizes the refined visual and text features by using the deformable encoder and transformer encoder, so that the hybrid queries can make full use of the contextual information of the referred object in the fusion decoding process. Ultimately, we generated object embeddings with better prediction accuracy while fully utilizing multimodal features.

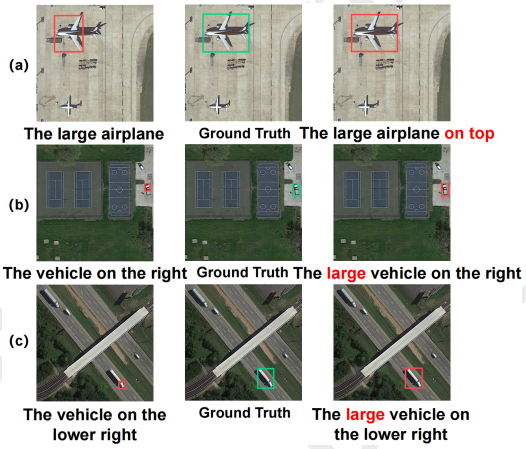

Figure 5: Analysis of Fault Case.

4.5 Fault Analysis

As shown in Figure 5, we observed that when multiple objects of the same category appear in an image, inadequate descriptions can lead to incomplete localization. For instance, there are two airplanes in Figure 5 (a). When the instruction is ”the large airplane”, the model only partially localizes the ground truth. But when we add more detailed location qualifiers like ”on top”, the model locates almost all the ground truth. Similar results can be seen in Figures 5 (b) and (c). When objects of the same category as the target but different sizes appear in

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

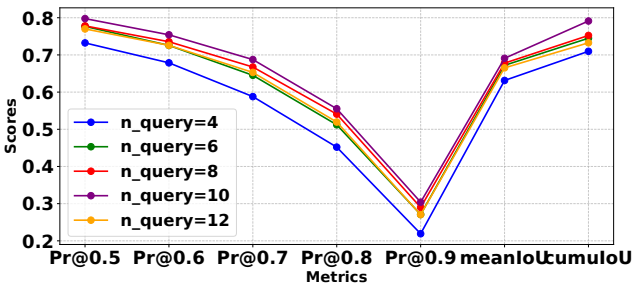

Figure 6: Analysis of Hybrid query numbers.

the same figure, the lack of sufficient location qualifiers will result in incomplete localization. However, complete localization is obtained after adding detailed qualifiers. This suggests that failures may arise when images contain objects of the same category but different scales. And this finding highlights the need to address this limitation in our future work.

4.6 Ablation Study

Hybrid query number: Our proposed LGFormer takes hybrid queries as the basis for decoding queries. The hybrid queries consist of sentence-level text representation and filtered visual representation, incorporating both textual and visual information. This provides rich prior knowledge for the initial decoding queries, enabling the model to learn more accurate localization information. To explore the impact of the number of hybrid queries on the model performance, we pre-train the model for five epochs on the DIOR-RSVG dataset and increase the number of hybrid queries from 4 to 12, respectively. As shown in Figure 6, we can see that the model performs best when the number of queries increases to 10. After that, as the number of queries continues to increase, the model's performance actually declines.

5 Conclusion

In this paper, we propose a language-guided hybrid representation learning framework called LGFormer. This structure uses hybrid queries we designed, which allow the queries to possess rich prior information during the decoding process, guiding themselves in the decoding process to extract multimodal contextual information. To better construct hybrid queries, we used sentence-level features and visual features obtained through the proposed language-guided visual feature filtering method. Finally, extensive experiments have shown that our model has achieved state-of-the-art performance on the DIOR-RSVG and OPT-RSVG datasets. However, there is still room for improvement in our model's performance, and our current method only allows for the localization of one object at a time. In the future, we will explore the visual grounding of complex objects in remote sensing images, while considering collaborative learning strategies for multi-source images.

Acknowledgments

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China (U22B2054), the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence(No. HMHAI-202404 and HMHAI-202405), the National Natural Science Foundation of China (62231027, 62431020, and 62471385), the 111 Project.

References

[Ashish, 2017] Vaswani Ashish. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017.

[Bajaj *et al.*, 2019] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. G3raphground: Graph-based language grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4281–4290, 2019.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[Deng *et al.*, 2021] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[Deng *et al.*, 2023] Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[Dogan *et al.*, 2019] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019.

[Dosovitskiy *et al.*, 2021] 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 *International Conference on Learning Representations*, 2021.

[Girshick *et al.*, 2015] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

IEEE transactions on pattern analysis and machine intelligence, 38(1):142–158, 2015.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2016] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016.

[Lan *et al.*, 2024] Meng Lan, Fu Rong, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Language query based transformer with multi-scale cross-modal alignment for visual grounding on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Li *et al.*, 2020] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

[Li *et al.*, 2024] Ke Li, Di Wang, Haojie Xu, Haodi Zhong, and Cong Wang. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Liao *et al.*, 2020] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[Nagaraja *et al.*, 2016] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Redmon, 2016] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[Rezatofighi *et al.*, 2019] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[Sun *et al.*, 2022] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 404–412, 2022.

[Wang *et al.*, 2016] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[Wang *et al.*, 2019] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019.

[Yang *et al.*, 2019a] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019.

[Yang *et al.*, 2019b] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4683–4693, 2019.

[Yang *et al.*, 2020a] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961, 2020.

[Yang *et al.*, 2020b] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020.

[Yang *et al.*, 2022] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508, 2022.

[Ye *et al.*, 2022] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15502–15512, 2022.

[Yu *et al.*, 2018] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[Zhan *et al.*, 2023] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[Zhu *et al.*, 2020] 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.