Omni-Q: Omni-Directional Scene Understanding for Unsupervised Visual Grounding

Sai Wang, Yutian Lin*, Yu Wu*
School of Computer Science, Wuhan University
{wangsai23, yutian.lin, wuyucs}@whu.edu.cn

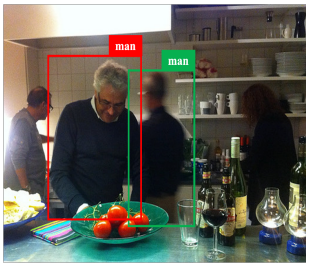
Abstract

Unsupervised visual grounding methods alleviate the issue of expensive manual annotation of image-query pairs by generating pseudo-queries. However, existing methods are prone to confusing the spatial relationships between objects and rely on designing complex prompt modules to generate query texts, which severely impedes the ability to generate accurate and comprehensive queries due to ambiguous spatial relationships and manually-defined fixed templates. To tackle these challenges, we propose a omni-directional language query generation approach for unsupervised visual grounding named Omni-Q. Specifically, we develop a 3D spatial relation module to extend the 2D spatial representation to 3D, thereby utilizing 3D location information to accurately determine the spatial position among objects. Besides, we introduce a spatial graph module, leveraging the power of graph structures to establish accurate and diverse object relationships and thus enhancing the flexibility of query generation. Extensive experiments on five public benchmark datasets demonstrate that our method significantly outperforms existing state-of-the-art unsupervised methods by up to 16.17%. In addition, when applied in the supervised setting, our method can freely save up to 60% human annotations without a loss of performance.

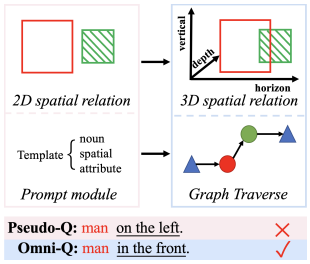
1. Introduction

Visual grounding [11, 15], also known as Referring Expression Comprehension (REC), has been rapidly developed in recent years. Its purpose is to locate relevant objects in an image based on a language query and serves as a fundamental component for various multi-modal tasks, such as image captioning [10, 41] and visual question answering [2, 27, 47].

The mainstream visual grounding methods heavily rely on manual-annotated dataset, which can be broadly categorized into two types: fully supervised [5, 9] and weakly-supervised [4, 32]. Fully supervised methods utilize region-query pairs provided in the dataset for training, while weakly-supervised methods solely rely on images and queries without any box information. Since clearly describing the appearance and location of the object is time-consuming and labor-intensive, annotating the language query becomes the bottleneck of dataset labeling.



(a) original image



(b) query of different methods

Figure 1. Comparison of queries generated by Pseudo-Q and **Omni-Q**. (a) The object locations with their descriptions. (b) The generation mechanism of Pseudo-Q and Omni-Q are highlighted by pink and blue boxes, respectively.

To alleviate the high annotation cost, some works attempt to address unsupervised REC without queries nor object boxes. Pseudo-Q [13] is one of the most representative methods, which leverages off-the-shelf detector and attribute classifiers to describe the object, and generate pseudo queries based on prompts. Since Pseudo-Q only captures information within the same category, it lacks the ability to model the inter-category object relationships. It also designs a complex prompt module to refine generated pseudo queries to tailor for visual grounding task.

However, the aforementioned method only models the spatial relationships according to the 2D coordinate values, *e.g.*, using the relative spatial relations of 2D pixels to represent their 3D relations in the real world, which may lead to wrong descriptions. As shown in Figure 1(a), the object described in the red box should rely on front-rear spatial re-

*Corresponding author.

lationship. However, Pseudo-Q relies solely on 2D spatial relation for comparison, and thus it is prone to generate ambiguous descriptions for the green box according to spatial position (Figure 1(b)). Additionally, it employs complex prompt templates or mechanisms to organize the object information into language queries. Although this approach seems to offer a variety of combination patterns, the manually designed templates ultimately limit the flexibility of query generation. The generated queries are consistently constrained to a predetermined set of fixed sentence structures, thereby impeding the generated queries in terms of their diversity and comprehensiveness. Consequently, some valuable information may be missed in the generated query.

In order to accurately model the spatial relationship between objects and flexibly generate comprehensive and diverse language queries, we propose a omni-directional language query generation (**Omni-Q**) approach, which consists of three modules: (1) The object perception module directly extracts object locations and rich descriptions, obviating the need to match the results of the object detector and attribute classifier in previous works. (2) The 3D spatial relation module innovatively extends the 2D spatial representation to 3D by decoupling relative and absolute positions. This allows the model to comprehend objects from both local and global perspectives, gaining insight into where the object is in the image and which object is nearby. (3) The spatial graph module builds upon the learned object semantic and spatial information, leveraging a graph structure to establish accurate and comprehensive spatial connections between objects, the capability to model spatial relationships which is significantly better than that of a scene graph. It achieves a seamless integration of objects and spatial relationships, without the need for sophisticated prompt mechanism to organize queries. By leveraging the synergistic effects of these three modules, Omni-Q is capable of flexibly generating precise and comprehensive language queries.

Extensive experiments on five datasets demonstrate the effectiveness of our method, which significantly outperforms all existing weakly-supervised and unsupervised visual grounding methods by up to 22.75% and 16.17%, respectively. Particularly, Omni-Q obtained competitive and even superior performance compared to fully supervised methods on certain evaluation metrics. Furthermore, when compete with supervised models, Omni-Q could achieve the same performance with only 40% of labeled data while the other 60% data kept unlabeled. It proves our method may potentially save up to an impressive 60% manual annotation costs without a loss of performance.

Our contributions are summarized as follows:

- We introduce a novel unsupervised visual grounding framework named Omni-Q, which extends spatial relationships into 3D for the first time and significantly enhances the accuracy of query generation.
- We propose a spatial graph module to utilize the graph structure to associate objects and spatial relationships, eliminating the need for complex prompt modules while maintaining the flexibility to generate diverse queries.
- Experiments shows that our method not only achieved state-of-the-art performance but also attained competitive results compared to fully supervised methods.

2. Related Work

2.1. Visual Grounding

Visual grounding [17, 48] is an important part of bridging language [6, 20] and visual representation [8, 31]. Existing visual grounding methods mainly divided into full-supervised [12, 14, 21, 30, 38], weakly-supervised [7, 34, 36, 46] and unsupervised [13, 33]. Fully supervised methods rely heavily on box-query pairs in annotations, which are expensive and time-consuming to annotate despite higher performance. Weakly-supervised methods try to alleviate this problem by exploiting image-query pairs, which obtain object proposals to match the query through an additional off-the-shelf detector. Although the labeling burden is reduced to a certain extent, the bottleneck mainly depends on the language queries. The unsupervised method does not need any labeled data and can generate box-query pairs only through image information, avoiding the problem of requiring a large amount of labeled data.

2.2. Unsupervised Visual Grounding

Among the existing unsupervised visual grounding methods [33, 40], the most representative work is Pseudo-Q [13], which directly generates box-query pairs from images. Pseudo-Q uses the object detector and attribute detector to obtain the location and attribute of the object, then determines the spatial relationship between same category objects through heuristic rules, and finally sends it to the query prompt module to obtain the query following templates. Although Pseudo-Q takes into account the category, attribute, and spatial relationship for intra-category objects, it does not model the inter-category object relationships, which restricts the ability to describe objects. At the same time, a large number of manually designed rules make it difficult to guarantee the accuracy of spatial relationships.

3. Proposed Method

3.1. Overview

Previous approaches mainly determine the spatial relationship on 2D plane and heavily relied on heuristic algorithms, employing a large number of manually designed rules to generate queries, which often hindered guaranteeing the correctness and comprehensiveness of queries. In this paper, we propose a novel unsupervised visual grounding

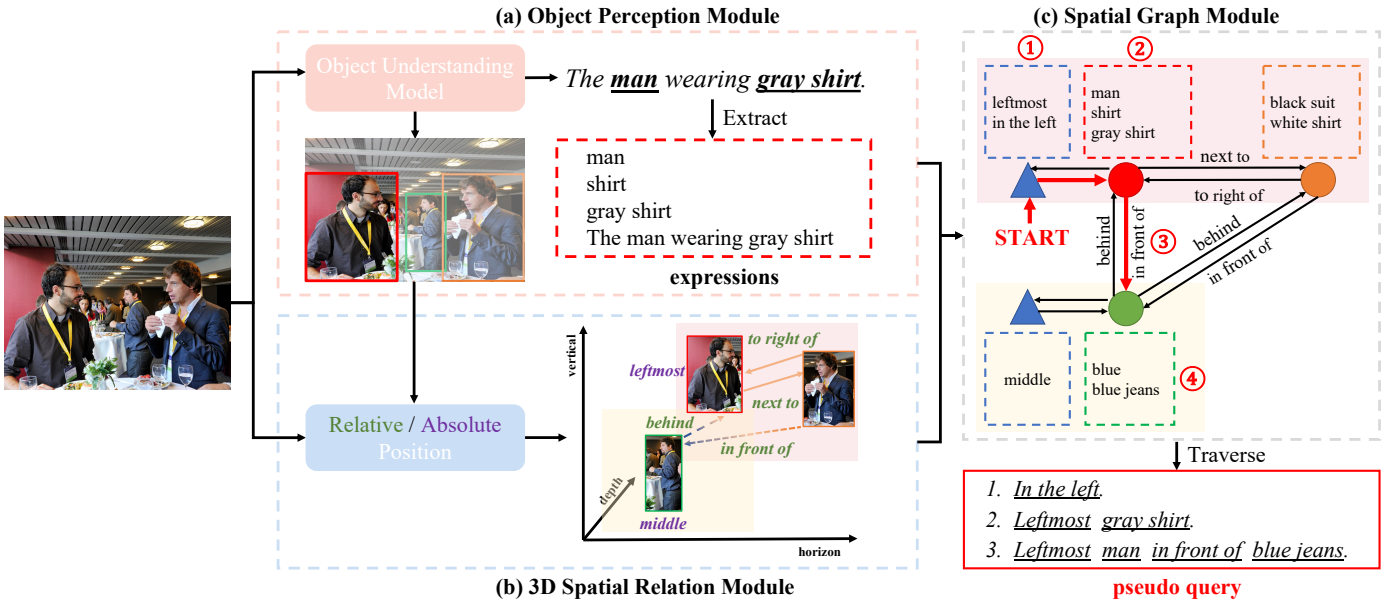


Figure 2. Overview of our method. We take unlabeled images as input and employ the **(a) object perception module** to extract the position and description of the objects. Afterward, position information is fed into the **(b) 3D spatial relation module** to produce the spatial relationships among the objects. Finally, we utilize the **(c) spatial graph module** to construct a spatial graph and traverse it to obtain the object queries.

framework, named **Omni-Q**, to generate correct and high-quality language queries. The overview of Omni-Q is depicted in Figure 2. Specifically, the object perception module takes an unlabeled image as input to extract the location and description of each object in the image. Subsequently, the image with object proposals are sent to the 3D spatial relation module, which produces the spatial relationship in three-dimensional representation based on depth information and plane position. By gathering the above elements, we obtain the spatial graph through the spatial graph module and traverse it to acquire region-query pairs enriched with semantic information for supervised learning.

3.2. Object Perception Module

The object perception module aims to extract both the location and rich descriptions of objects in images. Previous works [13] employ object detector and attribute classifier to generate object locations and descriptions, and they establish correspondences between objects and attributes using sophisticated matching algorithms, like matching clothes to people. However, the overall process becomes complex and lacks the high-level semantic required for a comprehensive description. In light of these considerations, we leverage the object understanding model trained on the same dataset, similar to previous work [13], to describe objects in a more natural way. This approach aims to incorporate higher-level semantic context into the object representation, allowing for a more complete and expressive description.

Specifically, the unlabeled image I is sent to the object understanding model [35] to obtain the bounding boxes b_i and descriptions s_i respectively. Although s_i tends to be more natural language style, it only describes the characteristics of a single object, which can lead to ambiguity among all elements and lacks global referentiality. To enhance the discriminativeness and diversity of s_i, we employ grammatical analysis to further break it down. Firstly, we use TextBlob [26] to parse the description s_i and obtain the syntax tree T_i:

$$T_i = TextBlob(s_i) \tag{1}$$

$$= \{\text{noun phrase}, \text{verb phrase}, \ldots, \text{prep}\}\,. \tag{2}$$

Then, we extract all noun phrase in T_i as expression candidate sets. In order to exclude irrelevant noun phrases and ensure the correctness of the generated query as much as possible, we retain only the first noun phrase in the category list. Next, we further process the noun phrase and extract adjectives such as color words to enrich the diversity of expression:

$$\{\text{object},\ \text{color},\ \text{cloth}\} = extract(\text{noun phrase}), \tag{3}$$

$$exp_i = s_i \cup \{\text{object},\ \text{color},\ \text{cloth}\}\,. \tag{4}$$

Finally, we aggregate the expressions of all objects in the image, and apply filter mechanism to retain non-duplicated

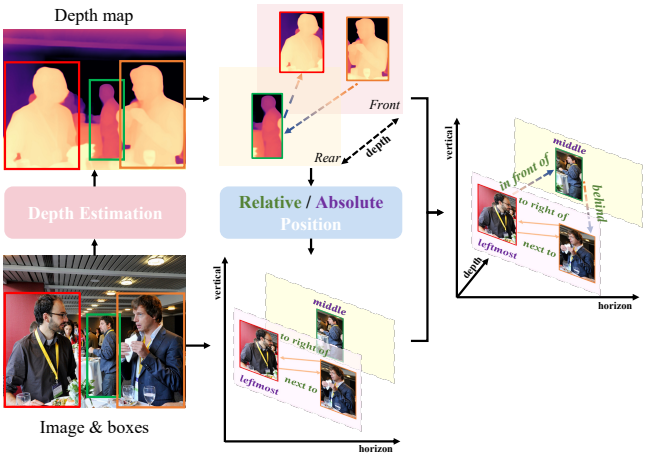


Figure 3. The architecture of 3D spatial relation module.

and unambiguous expressions $exp_i^{'}$:

$$\left\{exp_1^{'}, \cdots, exp_n^{'}\right\} = nondup(\{exp_1 \cup \cdots \cup exp_n\}), \tag{5}$$

where $nondup()$ denotes the filtering of unique elements. So far, we have obtained the object description pairs $\left\{(b_i, exp_i^{'})\right\}_{i=1}^{n}$ for subsequent graph construction. As shown in Fig. 2 (a), for the red bounding box, the corresponding $exp_i^{'}$ is generated with multiple description candidates in it.

3.3. 3D Spatial Relation Module

In this subsection, we propose the 3D spatial relation module to assemble spatial relationship for objects. Spatial relationships can be broadly categorized into three types: horizontal, vertical and depth. By leveraging 2D spatial relationships within the same depth plane along with attribute expressions, we can uniquely locate a specific object in three-dimensional representation. Moreover, we divide spatial relationships into relative positions and absolute positions, both of which are frequently used. The relative position indicates the positional relationship between the current object and other surrounding objects, which can describe object flexibly. The absolute position indicates the region in the image where the object is located, providing a simple and direct expression.

To model the 3D spatial relationship, we utilize a depth estimation model to generate the depth map D for the whole image. Then, the depth intensity of each object is calculated and compared with a threshold to determine the *fore-rear relationship*. Consequently, each object b_i is associated with a depth spatial information f_i, dividing the object into either foreground or background. Then, for objects of the same depth, we compare their position along both axial directions to obtain horizontal and vertical spatial information. Finally the relative positions Pos_r and absolute positions Pos_a are obtained by combining the horizontal, vertical and depth relationships:

$$\left\{pos_a^1, \cdots, pos_a^n\right\} = absolute(I, F, b_1, \cdots, b_n), \tag{6}$$

$$\left\{pos_r^{1,1}, \cdots, pos_r^{n,n}\right\} = relative(I, F, b_1, \cdots, b_n), \tag{7}$$

where pos_a^i denotes the absolute position of the i-th object, and $pos_r^{i,j}$ represents the relative position relationship between the i-th and j-th object.

The generated absolute and relative positions depict a comprehensive three-dimensional spatial representation for any object. Especially, challenging relationships such as front/rear and adjacent, which were prone to confusion in previous works can be easily addressed by our method.

3.4. Spatial Graph Module

In this subsection, we combine the object description and spatial relationship to construct a spatial graph, which allows us to flexibly use both to describe an object. With the spatial graph, we can establish communication between objects through adjacent edges, so that the current object can be described more diversely by other spatially adjacent objects. Note that our spatial graph differs from traditional scene graph, mainly in that a single node in our graph corresponds to multiple values, and the spatial position serves as the relationship between objects.

Concretely, we regard object expression $exp_i^{'}$ as a vertex in the graph, and the relationships $pos_r^{i,j}$ between i-th object and j-th object are represented as edges, resulting in a directed spatial graph denoted as G. As the absolute position pos_a^i of an object is only relevant to itself, it is also connected to the object as a distinct vertex. In graph G, each vertex corresponds to multiple descriptions of the objects, and these vertices are interconnected by directed edges, where the edge weight represents the spatial relationship between objects.

After the graph is built, queries can be generated by performing random walks among all the nodes. To start the process, we first determine the starting node and then randomly select one of its corresponding object descriptions as the value of the node. Starting from this initial node, we proceed by choosing a directed edge arbitrarily and moving to the next node in the graph. As we reach a new node, the sequence path from the starting node to the current node forms a query, which is added to the candidate set for subsequent model training. To generate diverse queries and ensure a comprehensive description of a certain object, we repeat the above process N times (i.e., visiting a maximum of N nodes), which yields traversal paths with lengths ranging from 1 to N, so as to achieve a differentiated description of

Algorithm 1: Spatial Graph Generation

Data: Object expressions $exp^{'} = \{exp_i^{'}\}_{i=1}^{N}$,
relative positions $Pos_r = \{pos_r^i\}_{i=1}^{N}$,
absolute positions $Pos_a = \{pos_a^{i,j}\}_{i,j=1}^{N}$
Result: Spatial graph G, candidate set Ω

Initialize empty spatial graph G and candidate set Ω;
foreach *index* i *in range(length(*$exp^{'}$*))* **do**
  Add $exp_i^{'}$ and pos_a^i as vertex to G;
  foreach *index* j *in range(length(*$exp_i^{'}$*))* **do**
    Add $(i, j, pos_r^{i,j})$ as directed edge to G;
while *epoch* $<$ *epochs* **do**
  Initialize empty string $query$;
  Randomly choose vertex $G[i]$ as current node;
  while *length of* $query < N$ **do**
    Add expression of $G[i]$ to $query$;
    Add $query$ to Ω;
    if *Node* $G[i]$ *has adjacent node* $G[j]$ **then**
      Add edge $G[i][j]$ to $query$;
      Set current node to $G[j]$;
    else
      break;
return spatial graph G and candidate set Ω;

a certain object, and generate accurate and diverse queries. The whole process of graph construction and query generation is illustrated in Algorithm 1.

4. Experiments

4.1. Datasets

Similar to previous method, we conducted experiments on five visual grounding datasets: RefCOCO [42], RefCOCO+ [42], RefCOCOg [28], ReferItGame [16] and Flickr30K Entities [29]. The number of training images in above datasets are 16994, 16992, 24698, 8994 and 29779 respectively. We use the same train/val/test splits for fair comparison and we DO NOT use any information of dataset except images for unsupervised training. We take the accuracy as the evaluation metric and the Jaccard overlaps between the predicted box and ground-truth larger than 0.5 is regraded as positive.

4.2. Implementation and Training Details

We adopt GRiT [35] as our object understanding model, which is pre-trained on Visual Genome [19] dataset same as Pseudo-Q. For a fair comparison, we use TransVG as our training model, which is end-to-end optimized with AdamW as optimizer and the initial learning rate is set to 2.5×10^{-5} for the visual and language encoder and 2.5×10^{-4} for the cross-modality fusion module. All our experiments are conducted with 8 Telsa V100 GPUs and the batch size is 256. Our model is trained for 30 epochs on all datasets and we use the same data augmentation strategy following TransVG [5]. For each image, we select 6 objects and uniformly sample up to 48 pseudo queries from candidates among all datasets, which aims to minimize the special design for the specific dataset.

4.3. Comparison with State-of-the-art Methods

RefCOCO/RefCOCO+/RefCOCOg. The performance of our method on RefCOCO, RefCOCO+ and RefCOCOg are presented in Table 1. Our method achieves significant improvement, surpassing the existing state-of-the-art method Pseudo-Q by a considerable margin. Additionally, our method achieves competitive performance when compared to the weakly-supervised state-of-the-art method DTWREG on all three datasets. Remarkably, it even outperformed some fully supervised methods on RefCOCOg. The results demonstrate that our method can generate high-quality queries with accurate spatial locations and comprehensive object relationships.

ReferItGame. As shown in Table 2, our method achieves an accuracy of 52.91% on ReferItGame dataset, and surpasses all state-of-the-art weakly-supervised and unsupervised methods. since the ReferItGame dataset contains a large number of spatial relationships, our method still achieves a steady improvement. In particular, our method further improves 14.52% and 9.59% compared to [34] and Pseduo-Q, which demonstrates the excellent ability of Omni-Q.

Flickr30K entities. As presented in Table 2, our method achieves a remarkable 65.23% top-1 accuracy on Flickr30K, which outperforms all the unsupervised as well as weakly-supervised methods. The dataset primarily consists of expressions related to object descriptions and attributes, with a minimal proportion of spatial relationships. Thanks to the flexibility of the graph structure, we set the maximum number of traversed nodes to 1, allowing the generated queries to focus on the object itself.

4.4. Ablation Studies

In this section, we conduct ablation experiments to verify the effectiveness of each component and thoroughly analyze the efficacy of our proposed method.

Effectiveness of object perception module. Table 3 shows that integrating the object perception module resulted in a performance improvement of 8.09% on RefCOCO. Despite the capability of object understanding model to directly generate descriptions for specific objects, its lack of uniqueness and spatial relationships prevents it from being suitable for object reference, which leads to significantly lower accuracy compared to other methods. Among the captions of all datasets generated by object understanding model, only 25.08% of them include spatial positions such

Method	Sup.	RefCOCOg			RefCOCO+			RefCOCO		
		val-g	val-u	test-u	val	testA	testB	val	testA	testB
MAttNet [43]	Full	-	66.58	67.27	65.33	71.62	56.02	76.65	81.14	69.99
NMTree [22]		64.62	65.87	66.44	66.46	72.02	57.52	76.41	81.21	70.09
FAOA [37]		56.12	61.33	60.36	56.81	60.23	49.60	72.54	74.35	68.50
ReSC [38]		63.12	67.30	67.20	63.59	68.36	56.81	77.63	80.45	72.30
TransVG [5]		66.56	67.66	67.44	63.50	68.15	55.63	80.32	82.67	78.12
VC [45]	Weak	33.79	-	-	-	34.60	31.58	-	33.29	30.13
ARN [23]		34.66	-	-	34.53	36.01	33.75	34.26	36.43	33.07
KPRN [24]		33.56	-	-	35.96	35.24	36.96	35.04	34.74	36.98
DTWREG [32]		43.24	-	-	39.18	40.10	38.08	39.21	41.14	37.72
CPT [39]	No	-	36.70	36.50	31.90	35.20	28.80	32.20	36.10	30.30
Pseudo-Q* [13]		49.82	46.25	47.44	38.88	45.06	32.13	56.02	58.25	54.13
Ours *		**65.99**	**63.62**	**62.36**	**46.22**	**51.71**	**38.60**	**67.39**	**71.64**	**61.94**
Δ		(+16.17)	(+17.37)	(+14.92)	(+7.34)	(+6.65)	(+6.47)	(+11.37)	(+13.39)	(+7.81)

Table 1. Comparison with state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg in terms of top-1 accuracy (%). "Sup." refers to supervision level. Our results are highlighted with red regions and surpass all weakly-supervised and unsupervised methods by a large margin. * denotes the model pre-trained on the Visual Genome [19] dataset.

Method	Sup.	ReferIt	Flickr30K
PIRC Net [18]	Full	59.13	72.83
Yu *et al.* [44]		63.00	73.30
FAOA [37]		60.67	68.71
ReSC [38]		64.60	69.28
TransVG [5]		69.76	78.47
KACNet [3]	Weak	33.67	46.61
MATN [46]		33.10	13.61
ARN [23]		26.19	-
Gupta *et al.* [7]		-	51.67
Liu *et al.* [25]		37.68	59.27
Wang *et al.* [34]		38.39	53.10
Yeh *et al.* [40]	No	36.93	20.91
Wang *et al.* [33]		26.48	50.49
Pseudo-Q* [13]		43.32	60.41
Ours*		**52.91**	**65.23**
Δ		(+9.59)	(+4.82)

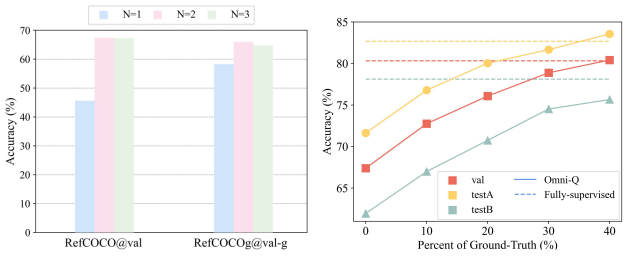
Table 2. Comparison with state-of-the-art methods on ReferItGame and Flickr30K Entities in terms of top-1 accuracy (%). "Sup." refers to supervision level. * denotes the model pre-trained on the Visual Genome [19] dataset.

as "on", "left", "front", etc. In contrast, the object perception module can extract key nouns and attributes from descriptions, thereby obtaining discriminative descriptions that contribute to the observed performance boost.

Effectiveness of 3D spatial relation module. As shown in Table 3, incorporating 3D spatial relation module (using Spatial+3D-Depth) greatly improves the performance of the model, with a maximum increase of 21.01% on RefCOCO. Absolute position and depth information provides clear spatial location for object expressions, enabling them to be explicitly referred to in 3D space, relying on both their own and surrounding objects position in the image.

OUM	OPM	Spatial	3D-Depth	SGM	RefCOCO
✓					37.42
✓	✓				45.51
✓	✓	✓			64.72
✓	✓	✓	✓		66.52
✓	✓	✓	✓	✓	67.39

Table 3. Ablations of each module on RefCOCO. "OUM" denotes only use descriptions generated by the object understanding model. "OPM" represents the object perception module, "Spatial" and "3D-Depth" belong to the 3d spatial relation module, which respectively denotes using spatial relation of position and depth information. "SGM" means the spatial graph module, specifically referring to the situation of visiting more than one node.

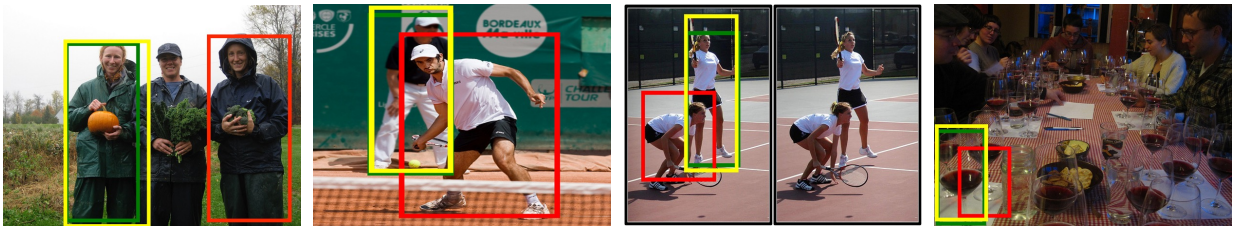


(a) (b)

Figure 4. (a) Ablation of the maximum visited nodes on RefCOCO and RefCOCOg. (b) Experiments on saving manual labeling on RefCOCO. The dashed line corresponds to the upper bound of the performance (i.e., the fully supervised accuracy).

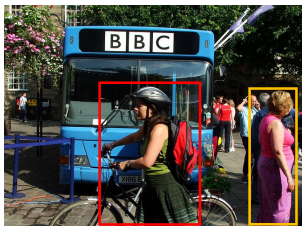
Effectiveness of spatial graph module. The spatial graph module was introduced to organize and generate queries by establishing relationships between objects. As shown in the last row of Table 3, further incorporating object relationship resulted in performance improvements of 0.87% on RefCOCO, respectively. This improvement is at-

guy with black bike shirts on near front | in back second person from left | man behind the child | white laptop in front of guy with dark shirt

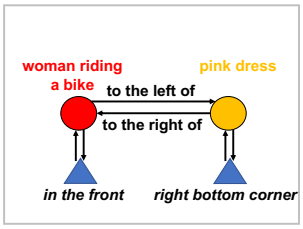


woman with pumpkin | guy in black shirt | person in left picture with racquet in the air | lower left glass

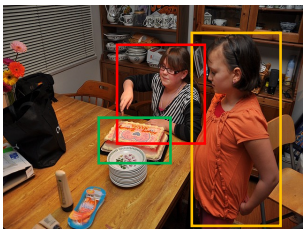
Figure 5. The visualization of detection results on RefCOCO. The green boxes and the description below the images are ground-truth. The yellow boxes and the red boxes correspond to the results of our method and Pseudo-Q.



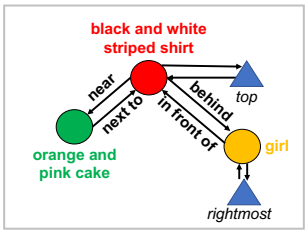
Pseudo-Q:
1. woman
2. left woman
3. woman skirt right
4. woman pink shirt on the right



Omni-Q:
1. woman riding a bike
2. in the front woman riding a bike
3. right bottom corner pink dress
4. pink dress right bottom corner to the right of woman riding a bike



Pseudo-Q:
1. shirt woman
2. girl orange shirt
3. cake



Omni-Q:
1. black and white striped shirt
2. rightmost girl in front of black and white striped shirt
3. orange and pink cake

Figure 6. Comparison of queries generated by Pseudo-Q and our method. The spatial graph corresponding to the image is shown on the right. For visual clarity, some elements and alternatives have been omitted.

Dataset		Pseudo-Q	**Ours**
RefCOCOg	val-g	39.56	**51.78** (+12.22)
	val-u	39.97	**51.51** (+11.54)
	test-u	40.02	**51.08** (+11.06)
RefCOCO+	val	29.23	**41.28** (+12.05)
	testA	34.51	**45.20** (+10.69)
	testB	23.30	**32.48** (+9.18)
RefCOCO	val	48.86	**57.61** (+8.75)
	testA	52.43	**60.79** (+8.36)
	testB	44.51	**48.40** (+3.89)

Table 4. Comparison of generalization performance (ReferIt → RefCOCO series) between Pseudo-Q and Omni-Q.

tributed to the ability of spatial graph module to establish relationships between various objects, and its flexibility in generating language queries using the graph structure.

Generalization study. To demonstrate the generalization ability of our method, we train the model using queries generated on ReferItGame and test it on RefCOCO series datasets, as shown in Table 4. We observe that compared to Pseudo-Q, Omni-Q exhibits the best generalization performance across all datasets. This finding suggests that our approach is not specifically designed for a particular dataset but rather a general unsupervised visual grounding framework built on a thorough understanding of images.

Number of maximum visited nodes. We conduct ablation experiments to explore the optimal number of maximum visited nodes during the traversal of the spatial graph. Increasing the number of visited nodes indicates that the generated queries include a greater number of objects and

Model	Training data	RefCOCO+		
		val	testA	testB
Pseudo-Q	P	38.88	45.06	32.13
	P + O	40.80	44.41	36.59
Omni-Q	P	43.26	50.77	34.36
	O	46.22	51.71	38.60

Table 5. Comparison of Pseudo-Q using different training data. "P" and "O" denote the data generated by attribute classifier [1] and object understanding model [35], respectively.

encompass more spatial relationships. As observed in Figure 4(a), with the increase in the maximum number of visited nodes, the overall accuracy shows an initial rise followed by a plateau. This suggests that involving more objects in the query may also introduce redundant information. Hence, we opt to traverse only two nodes.

The proportion of manual labeling saved. In Figure 4(b), we validate the effectiveness of generated queries by adding a certain proportion of ground-truth. Meanwhile, we use the accuracy achieved by the same model under fully supervised setting as the upper-bound. Our goal is to explore the minimum amount of ground truth required for the queries to enable the model to reach the upper bound performance. As a result, Omni-Q can effectively replace the vast majority of queries (i.e., approximately 60% ground truth) and achieve results comparable to fully supervised methods, significantly reducing the cost of manual annotation, far exceeding the 31% labeling savings of Pseudo-Q.

Study on object understanding model. To further study the performance gains brought by the object understanding model, we additionally input all the boxes and object descriptions generated by the same object understanding model to the training set of Pseudo-Q. In this way, Pseudo-Q would have more data information sources than our Omni-Q, i.e., the attribute classifier [1] (denoted as "P") and the object understanding model [35] (denoted as "O"). Both of P and O are pre-trained on the Visual Genome dataset. In addition, we also reproduce our model using exactly the same training data (P) with Pseudo-Q. Results are shown in Table 5. We can see object understanding model also help the training of Pseudo-Q, improving its validation accuracy from 38.88 to 40.80. But we can also see even though Pseudo-Q uses more training data (P + O) than our Omni-Q, its performance is still lower than Omni-Q. In addition, we can see when using identical training data P, Omni-Q achieves an improvement of over 5% compared to Pseudo-Q. This implies that simply integrating object descriptions is insufficient for a significant enhancement of the model. It is crucial to rectify spatial relationship errors during the query generation process and establish interconnections between objects to achieve a noticeable improvement in performance, aligning with our 3D spatial relation module and spatial graph module.

4.5. Qualitative Results

In this section, we visualize the detection results of the model and compare the queries generated by different methods to analyze the quality of the queries.

Visualization of detection results. To demonstrate the performance of our method, we present the visual results of the RefCOCO test dataset in Figure 5. The first row focuses on exploring the accuracy of model in handling depth-queries, while the last row assesses the ability of model to discriminate between similar and confusing objects. From the results, it is evident that three-dimensional spatial representation enables the model to accurately distinguish the ordering of objects. Furthermore, our method can effectively discern the composition relationships in queries, accurately identify the central words in queries, and locate the regions in the graph where they refer to.

The quality of generated queries. In addition, we display the generated queries in Figure 6. As depicted in the figure, the queries generated by Omni-Q are more accurate, maintaining precise spatial relationships while having rich descriptions. Notably, it overcomes the challenge of previous methods misidentifying "front/rear" as "left/right". Furthermore, the spatial graph effectively associates objects using spatial relationships, which affords us the opportunity to generate diverse queries flexibly. We calculate the average BERT feature similarity of quires generated by Pseudo-Q (0.81) and ours (0.71). The results show our generated queries have more information and less redundancy.

5. Discussions and Conclusion

Limitation. In the unsupervised setting, assessing the accuracy of the input source is challenging, and there are inevitably incorrect object expressions. Pseudo-Q similarly encounters this challenge, and enhancing performance in the future involves filtering high-quality queries. Secondly, exploring how to delve deeper into the analysis of image content is essential for future study.

Conclusion. We introduced Omni-Q to address the issues of confusing spatial relationships and the need for complex prompt modules in existing methods. Specifically, we proposed the 3D spatial relation module to extend 2D spatial representations to 3D to accurately localize the spatial location of objects. Additionally, the spatial graph module organizes objects leveraging graph structures and generates flexible and comprehensive queries. Extensive experiments demonstrate the effectiveness of our method, which significantly surpasses the state-of-the-art methods while saving nearly 60% of manual annotations.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 62372341.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 8

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018. 6

[4] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019. 1

[5] 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. 1, 5, 6

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[7] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2, 6

[8] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[9] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 684–696, 2019. 1

[10] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1

[11] 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. 1

[12] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124, 2017. 2

[13] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022. 1, 2, 3, 6

[14] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[15] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014. 1

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5

[17] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2

[18] Rama Kovvuri and Ram Nevatia. Pirc net: Using proposal indexing, relationships and context for phrase grounding. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 451–467. Springer, 2019. 6

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 5, 6

[20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 2

[21] 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. 2

[22] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 6

[23] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019. 6

[24] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression

grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019. 6

[25] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5612–5621, 2021. 6

[26] Steven Loria et al. textblob documentation. *Release 0.15*, 2 (8):269, 2018. 3

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 1

[28] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5

[30] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. In *European Conference on Computer Vision*, pages 546–562. Springer, 2022. 2

[31] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[32] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4189–4195, 2021. 1, 6

[33] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4672, 2019. 2, 6

[34] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021. 2, 5, 6

[35] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 3, 5, 8

[36] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 2

[37] 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. 6

[38] 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. 2, 6

[39] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 6

[40] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6125–6134, 2018. 2, 6

[41] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 1

[42] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5

[43] 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. 6

[44] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv preprint arXiv:1805.03508*, 2018. 6

[45] Hanwang Zhang, Yulei Niu, Tianlang Chang, Shih-Fu, Wengang Zhou, and Houqiang Li. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 6

[46] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018. 2, 6

[47] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 1

[48] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+ x: A survey on multimodal learning in the light of data. *arXiv preprint arXiv:2210.02884*, 2022. 2