CLIPREC: Graph-based Domain Adaptive Network for Zero-Shot Referring Expression Comprehension

Jingcheng Ke, Jia Wang, Jun-Cheng Chen, *Member, IEEE,* I-Hong Jhuo, *Member, IEEE,* Chia-Wen Lin, *Fellow, IEEE,* and Yen-Yu Lin, *Senior Member, IEEE*

***Abstract*—Referring expression comprehension (REC) is a cross-modal matching task that aims to localize the target object in an image specified by a text description. Most existing approaches for this task focus on identifying only objects whose categories are covered by training data. This restricts their generalization to unseen categories and practical usage. To address this issue, we propose a domain adaptive network called CLIPREC for zero-shot REC, which integrates the Contrastive Language-Image Pretraining (CLIP) model for graph-based REC. The proposed CLIPREC is composed of a graph collaborative attention module with two directed graphs: one for objects in an image and the other for their corresponding categorical labels. To carry out zero-shot REC, we leverage the strong common image-text feature space from the CLIP model to correlate the two graphs. Furthermore, a multilayer perceptron is introduced to enable feature alignment so that the CLIP model is adapted to the expression representation from the language parser, resulting in effective reasoning from expressions involving both seen and unseen object categories. Extensive experimental and ablation results on several widely-adopted benchmarks show that the proposed approach performs favorably against state-of-the-art approaches for zero-shot REC.**

***Index Terms*—Referring expression comprehension, domain adaptive network, zero-shot learning, CLIP**

I. Introduction

BRIDGING human beings and machines for efficient human-computer interaction in real-world applications has become a research hotspot of artificial intelligence [1]. One promising direction to realize this goal is through referring expression comprehension (REC) [2], [3], [4], which enables machines to comprehend the natural language in text from humans and find out the target object in an image. The core of this task is multi-modal reasoning between the textual expression semantics and the visual target objects. For example, for a given expression "a dog on the left of a cat", this task requires a machine to comprehend the expression and identify the object "cat" in a given image, followed by locating the dog indicated by the direction noun "left" and the object "cat" in the image. Since object category and expression annotations in our daily lives are extensive, it is almost infeasible to collect and label all possible image-expression pairs manually. Thus, zero-shot REC, as shown in Fig. 1, is more practical in the real-world setting than the standard REC. However, it is a more challenging task due to the requirement of figuring out the underlying relationships between unseen target objects and given expressions.

Most existing REC methods [5], [6], [7], [8], [9] are supervised. Unsupervised or zero-shot REC, *e.g.*, [10], was rarely explored in the literature due to its difficulties. Existing approaches in zero-shot REC tend to primarily focus on local content, resulting in inferior results when complex expressions and global reasoning are required. On the other hand, Wang *et al*. [11] proposed an unsupervised REC method requiring neither training nor paired phrase localization annotations. The method localizes the target object by leveraging semantic similarities between the query expression and the predicted concept labels of object candidates from various visual detectors. However, this method does not utilize the visual features of object candidates for matching. It thus results in suboptimal performance when the detected concepts are insufficient to differentiate the appearances of the target from other similar objects in the image. In this work, we aim to handle unseen objects during the inference time and focus on zero-shot REC by leveraging the pre-trained CLIP model [12]. The CLIP model has shown superior performance on zero-shot image classification because it is pre-trained contrastively using 400 million image-text pairs. To better use the power of CLIP, we adapt it to REC by using the training data of the target dataset. Specifically, we propose a novel graph-based domain adaptive network for zero-shot REC called CLIPREC, that integrates the Contrastive Language-Image Pretraining (CLIP) model [12] into a two-stage graph-based algorithm. Graph structures can exploit both local and global context information from detected objects, and is flexible to work with additional modules. Through the common feature space learned by CLIP, our graph-based method can infer the

Manuscript received November 20, 2022; revised April 7, 2023 and May 30, 2023; accepted July 7, 2023. Date of publication month day, 2023; date of current version month day, 2023. This work was funded in part by National Science and Technology Council (NSTC) under grants 111-2221-E-004-010, 111-2628-E-A49-025-MY3, 112-2221-E-A49-090-MY3, 111-2634-F-002-023, 111-2634-F-002-022 and in part by Qualcomm Technologies, Inc., through a Taiwan University Research Collaboration Project, under Grant NAT-487844. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhenzhen Hu. (Corresponding author: Jun-Cheng Chen)

J. Ke and C.-W. Lin are with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: freedom6927@gmail.com; cwlin@ee.nthu.edu.tw).

J. Wang is with the Department of Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan (e-mail: vicky.ee08@nycu.edu.tw).

J.-C. Chen is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115201, Taiwan (e-mail: pullpull@citi.sinica.edu.tw).

I.-H. Jhuo is with Microsoft, Seattle, WA, USA (e-mail: ihjhuo@gmail.com).

Y.-Y. Lin is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan (e-mail: lin@cs.nycu.edu.tw).

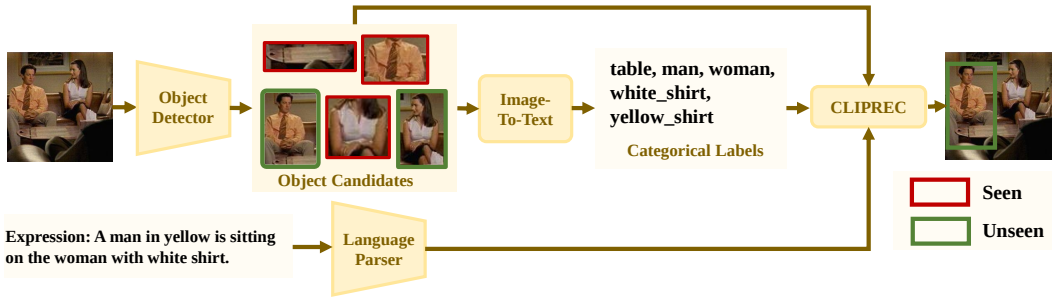


Fig. 1. Conceptual diagram of the proposed CLIPREC. It leverages the rich image-text information encoded in the CLIP (contrastive Language-Image Pretraining) model to overcome the difficulties of zero-shot REC (referring expression comprehension) where the target object categories specified by a given expression can be unseen during the training stage of the REC model.

corresponding categorical labels even for object candidates whose categories are not covered in the training data.

The CLIP model is mainly derived to extract features for image-text matching instead of reasoning and exploring the structure and semantics of the given images and expressions. Namely, it is not optimized for REC. Before using CLIP for zero-shot matching unseen objects, we propose performing feature alignment (FA) to synchronize the CLIP features and the expression representations obtained by the language parser. The FA module in this work uses trainable MLP (multilayer perceptron) over the CLIP features and projects all the features to a common subspace.

The proposed domain-adaptive network contains a graph collaborative attention module with two directed graphs, one for the objects of a given image and the other for the categorical labels of the objects. It leverages the CLIP features for working with unseen object categories and is developed upon GCN (graph convolutional networks) [13] for information aggregation. Two additional trainable multilayer perceptrons are appended after the image and text encoders of the pre-trained CLIP model for feature alignment. The resultant visual and textual features produced by the CLIP model can be better aligned with the expression representation from the language parser to improve performance further. Finally, the matching can be done by finding the pair with the highest similarity score between the expression and graph nodes.

In short, our contributions are two-fold: (1) we proposed a novel model called CLIPREC, which effectively tackles the challenging zero-shot REC problem by leveraging the CLIP model, pre-trained on a rich collection of image-text pairs. We utilize a graph-based multi-modal aggregation strategy to adapt CLIP for the downstream REC task, effectively aggregating CLIP features for visual objects and textual expressions. This strategy ensures that the CLIP features and expression representations obtained from language processing are synchronized, effectively bridging the domain gap. (2) our extensive experimental results demonstrate that our CLIPREC performs favorably against state-of-the-art methods for the target REC task with unseen object categories in a zero-shot manner.

II. Related Work

The literature about referring expression comprehension (REC) is quite extensive. In the following, we review the research topics highly relevant to this work, including zero-shot learning and graph-based REC.

A. Zero-Shot Learning with CLIP model

Zero-shot learning [14], [15], [16], [17], [18] has been studied as a way of generalizing a trained model to unseen object categories for different vision applications. Research efforts have been made to develop effective approaches to accomplish zero-shot learning. For example, Frome *et al*. [15] improved the method presented in [16] and demonstrated that after the trained model is fine-tuned on the ImageNet dataset [19]. Li *et al*. [17] enabled zero-shot transfer to several existing computer vision classification benchmarks via exploiting natural language supervision. By fine-tuning convolutional neural networks pre-trained on the ImageNet dataset, their approach achieves promising performance in predicting 30 million Flickr photos with a much wider set of visual concepts in a zero-shot manner.

OpenAI released the CLIP model [12], which leverages 400 million image-text pairs collected from the Internet to contrastively train the model and achieve powerful joint image-text representations. The CLIP model achieves promising results in the image classification task in the zero-shot setting.

Recently, some methods based on CLIP were proposed to solve zero-shot tasks. For example, Gu *et al*. [20] proposed an open-vocabulary object detector for zero-shot object detection by exploiting CLIP, which leads to satisfactory performance. Jia *et al*. [21] constructed a large-scale dataset by feeding 1.8B image-text pairs into the designed image-based and text-based filtering. The constructed dataset was used to train their proposed method, named ALIGN, which adopts contrastive learning to cluster matched and non-matched image-text pairs in each batch, respectively, for zero-shot classification. Recently, the method proposed in [22] describes the textual events of an expression by a series of prompt functions and identifies the object candidates of an image. The resultant descriptions and proposals are fed into the textual and visual encoders of CLIP for matching. Li *et al*. [23] proposed BLIP (Bootstrapping Language Image Pretraining) for extracting better image-text representation. BLIP adopts a multimodal mixture of encoder-decoder framework and leverages a captioner and a filter to remove the noisy captions to create a better large-scale image-text dataset for multi-task pre-training. BLIP improves

the performance of zero-shot text-to-video retrieval and zero-shot video question answering after fine-tuning the pre-trained BLIP for each downstream task. As reported in [24], CLIP lacks the sensitivity of spatial relationships of objects in images and is thus not optimal for reasoning. To address this issue, Subramanian *et al*. [24] proposed a zero-shot REC method named ReCLIP, which takes the object candidates of an image and noun chunks of an expression as input to the visual and textual encoders of CLIP, respectively, for matching. Although their method makes use of CLIP to facilitate zero-shot REC, the CLIP model is adopted without any adaptation and probably suffers from the domain gap between the data used to pre-train the CLIP model and the data used for REC.

Different from [24], our method integrates the CLIP model into the proposed graph-based domain adaptive network. With our CLIPREC, the CLIP features for visual objects and textual expressions can be aggregated via the graph structure for multi-modal reasoning and REC. Furthermore, we include a feature alignment module for the CLIP model so that training the network jointly can well adapt the CLIP model to the downstream REC task. It turns out that our method makes the most of both worlds of CLIP features for unseen objects and the language parser for expression analysis. The proposed method can also effectively exploit the correlation of the visual information of object candidates and their corresponding categorical information in the zero-shot setting by leveraging the rich, joint image-text representations after adapting the pre-trained CLIP model.

B. Graph-based REC

In existing REC methods, graph-based methods can more intuitively capture the relationship between objects in an image and are more suitable for language-guided reasoning. Reasoning is widely used in text-to-image tasks. For example, Liang *et al*. [25] proposed a new task by introducing abductive reasoning to computer vision. Existing methods for graph-based REC typically employ two-stage pipelines because their graphs are constructed based on object candidates. Some graph-based REC methods [6], [26], [27], [28], [29], [30] were proposed to learn the cross-modal correlation between the vision and language domains. For example, the methods proposed in Wang *et al*. [27] and Yang *et al*. [28] employ a language parser to analyze linguistic structures for complex expressions, and construct an attention graph by jointly referring to the textual features of expressions and the visual features of objects. Under the guidance of expressions, the object, *i.e*., a node in the graph, with the highest score, is considered as the target object instance that matches the expression. Different from the aforementioned methods, He *et al*. [26], Yang *et al*. [29], and Jing *et al*. [6] proposed to construct the scene graph based on an expression. The scene graph is then used to guide the construction of another graph, called the appearance graph, for the objects in an image. The similarity of the two graphs is computed and used to identify the target object described by the expression.

Although these graph-based methods have achieved promising performances, they mainly use the visual features of an image and textual features of expressions for matching, ignoring the categorical features of the objects. Therefore, the results can be further improved, especially when grounding using complex referring expressions. Different from existing approaches, we jointly exploit the visual features and categorical features of the objects in an image with our proposed graph-based domain adaptive network. We find that jointly exploiting both pieces of information helps improve the performance of REC. More importantly, while existing methods for REC are applicable to object categories covered by training data, our method can accomplish the grounding task even for unseen object categories in a zero-shot manner.

C. Text-guided object segmentation and detection

Text-guided video segmentation (TVS) aims to extract objects in a video based on a textual description. Liang *et al*.proposed two frameworks for TVS: ClawCraneNet [31] and Local-Global Context Aware Transformer (LOCATER) [32]. ClawCraneNet locates the target object among all candidate objects using the feature maps of video frames, guided by the positional relation among objects, textual relation of expressions, and inter-frame temporal relation of the video. Meanwhile, LOCATER proposes a memory consisting of two components: a global memory and a local memory. The global memory gathers those frames representing the global visual contents, and the local memory gathers diverse temporal contexts of video based on the last segmented frame and current local memory state. The encoded expression, along with each frame of video, frames in the global memory, and the local memory, are then fed into the referring decoder to mask the target object.

Text-guided object detection (TOD), a related task to TVS, also called open-vocabulary object detection, has attracted increasingly more interest within the community. Open-vocabulary object detection aims to locate unseen objects by the learned knowledge from some seen objects. In particular, Li *et al*. [33] presented the first work on open-vocabulary object detection with textual descriptions. Their proposed method encodes a series of noun-descriptions of unseen objects by an LSTM model and concatenates them to the RoI pooling features of objects from the detector for matching. Feng *et al*. [34] proposed a learnable prompt in front of the pre-trained text encoder to embed the categorical labels. Then, the embedded categorical labels are fed into the pre-trained text encoder to guide the detector for classification and bounding box regression. Gu *et al*. [24] proposed a knowledge distillation system that guides the detector to locate unseen objects in an image using CLIP's encoded object proposals and categorical labels.

Zero-shot REC and TOD share some similarities, but there are also important differences between them. Specifically, in zero-shot REC, the task involves an entire sentence that describes an unseen object, including both the noun phrases and the relationships between them. This requires analyzing the given sentence and identifying the object it describes in an image, rather than simply locating an object related to the subject of the sentence in the image, as in TOD. As a result, the constraints on zero-shot REC are stronger than those on TOD.

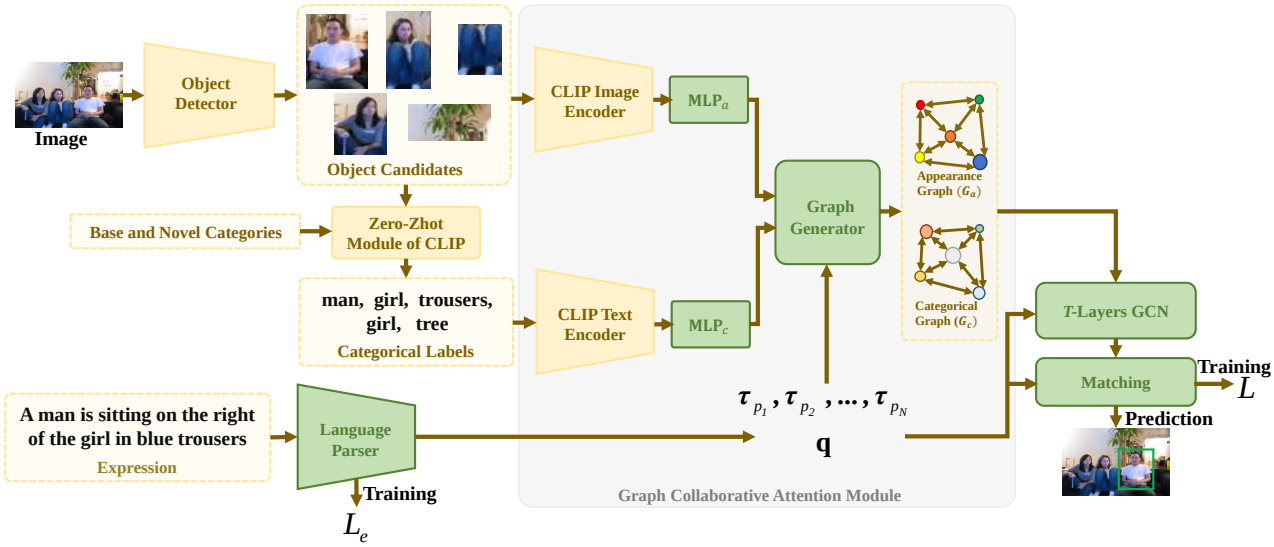


Fig. 2. Proposed CLIPREC framework for zero-shot REC. First, the expression is represented by $N+1$ representations, $\boldsymbol{\tau}_1 \ldots \boldsymbol{\tau}_N$ and $\mathbf{q}$, for N objects of expression and the entire expression, respectively. The pretrained CLIP model is then employed to infer the categorical labels of object candidates, where $N+1$ expression representations with the adapted CLIP features of the object candidates and their corresponding category labels are used to construct two attention graphs $\mathcal{G}_a$ and $\mathcal{G}_c$. In each graph, the node weights represent the relationship between the proposal and the expression, and the edge weights represent the relationship between the two connected nodes and the expression. Finally, the GCN-based multi-step reasoning is performed on the two attention graphs, followed by matching the target object in the image specified by the given expression. Larger node sizes in the graph denote larger assigned weights for each graph. Parameters in the green and yellow blocks are trainable and fixed, respectively.

III. Proposed Method

A. Overview

Fig. 2 illustrates the proposed CLIPREC framework that integrates the CLIP model into the graph-based domain adaptive network for zero-shot REC. CLIPREC consists of four modules: a language parsing module, a graph collaborative attention module, a multi-step reasoning module, and a matching module. The language parser encodes a given expression into N feature vectors to summarize the expression's N nouns (entities). The graph attention module constructs two attention graphs: one for objects in the input image and the other for their corresponding labels with the expression where the visual and categorical features are extracted using the pre-trained CLIP model appended with trainable, multilayer perceptrons for feature adaptation. The multi-step reasoning module consists of a T-layer GCN (graph convolutional network) aiming to aggregate information extracted from the two graphs, with which the matching module infers the target object with the highest similarity score between the given expression and the resultant node features of the two graphs.

The pseudo-code listed in **Algorithm 1** summarizes the proposed CLIPREC, whose details are elaborated in the following subsections. Compared with existing graph-based REC methods, the key novelties of CLIPREC lie in the following aspects. First, we devise a new self-attention strategy to represent the noun chunks of an expression to improve presentation accuracy (corresponding to Steps 6–7 in the pseudo-code). Second, we propose a multi-modal graph construction strategy (corresponding to Steps 8–11 in the pseudo-code) that utilizes the zero-shot module of CLIP, the representation of noun chunks, and a multilayer perceptron to construct both an attention appearance graph and a categorical graph. The two graphs are employed to better align the CLIP domain with the language parser.

Algorithm 1 CLIPREC

1: **Required symbol:** visual graph $\mathcal{G}_a$, categorical graph $\mathcal{G}_c$, visual graph after T reasoning steps $\mathcal{G}_a(T)$, categorical graph after T reasoning steps $\mathcal{G}_c(T)$
2: **Required variable:** words of expression $\{\mathbf{h}_\ell\}_{\ell=1}^{L}$, expression $\mathbf{q}$, noun-phases of expression $\{\boldsymbol{\tau}_n\}_{n=1}^{N}$, visual objects $\{\mathbf{o}_i\}_{i=1}^{K}$, positions of visual objects $\{\mathbf{b}_i\}_{i=1}^{K}$, visual object labels $\{\boldsymbol{\zeta}_i\}_{i=1}^{K}$, number of samples in each minibatch m
3: **Required function:** language parser $\mathbf{LP}$, graph construction $\mathbf{GC}$, mutli-step reasoning $\mathbf{MR}$, matching $\mathbf{M}$
4: **for** each minibatch do **do**
5: **for** $iter = 1$ to m **do**
6: Represent N nouns of expression (*i.e.*, (1)–(3))
7: $\{\boldsymbol{\tau}_n\}_{n=1}^{N} \leftarrow \mathbf{LP}\left(\{\mathbf{h}_\ell\}_{\ell=1}^{L}\right)$
8: Construct the appearance graph (*i.e.*, (4)–(10))
9: $\mathcal{G}_a \leftarrow \mathbf{GC}\left(\{\mathbf{o}_i\}_{i=1}^{K}, \{\mathbf{b}_i\}_{i=1}^{K}, \{\boldsymbol{\tau}_n\}_{n=1}^{N}\right)$
10: Construct the categorical graph (*i.e.*, (4)–(10))
11: $\mathcal{G}_c \leftarrow \mathbf{GC}\left(\{\boldsymbol{\zeta}_i\}_{i=1}^{K}, \{\mathbf{b}_i\}_{i=1}^{K}, \{\boldsymbol{\tau}_n\}_{n=1}^{N}\right)$
12: Multi-step reasoning (*i.e.*, (11)–(13))
13: $\mathcal{G}_a(T), \mathcal{G}_c(T) \leftarrow \mathbf{MR}(\mathcal{G}_a, \mathcal{G}_c, \mathbf{q})$
14: Matching (*i.e.*, (14)–(16))
15: prediction $\leftarrow \mathbf{M}(\mathcal{G}_a(T), \mathcal{G}_c(T), \mathbf{q})$
16: **end for**
17: **end for**

B. Language Parser

Due to the excellent performance of self-attention-based bi-direction LSTM (Bi-LSTM) [35] for language parsing [36],

[27], [37], [38], we employ a Bi-LSTM language parser with our proposed self-attention strategy for more precise attention. Given an expression Q of L words, each word is embedded into a vector by a non-linear mapping function, resulting in the embedded vectors $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_L\}$. The parser encodes the embedded vectors into a vector sequence $H = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_L\}$ by using Bi-LSTM, where $\mathbf{h}_\ell$ is the concatenation of the output of forward and backward Bi-LSTM at the ℓ-th word. Meanwhile, the overall representation of the expression is denoted by a feature vector $\mathbf{q}$, which is the concatenation of the last hidden states of both the forward and backward LSTMs.

Let N denote the number of nouns (entities) in the expression. We represent the text embedding of N nouns as $\{\mathbf{e}_{p_1}, \mathbf{e}_{p_2}, ..., \mathbf{e}_{p_N}\}$ where p_n, for $n = 1, ..., N$, is the index of the word corresponding to the n-th noun in the expression. For our self-attention strategy, we derive N groups of weights, one for each noun, to emphasize the object-related words in the expression. Specifically, for the n-th noun, the corresponding weight group $\alpha_n = \{\alpha_{n,\ell}\}_{\ell=1}^{L}$ consists of L weights for highlighting the words related to this noun. We optimize a learnable vector $\mathbf{w}_n$ to derive the weight group α_n by applying $\mathbf{w}_n$ to $H = \{\mathbf{h}_1, ..., \mathbf{h}_L\}$ followed by the softmax function, namely

$$\alpha_{n,\ell} = \frac{\exp(\mathbf{h}_\ell^T \mathbf{w}_n)}{\sum_{t=1}^{L} \exp(\mathbf{h}_t^T \mathbf{w}_n)}, \text{ for } 1 \leq \ell \leq L. \tag{1}$$

With the weight group α_n, the expression representation of the n-th noun can be represented by

$$\boldsymbol{\tau}_n = \sum_{\ell=1}^{L} \alpha_{n,\ell} \mathbf{h}_\ell, \tag{2}$$

which is a weighted combination of the word vectors with the relevant words emphasized. By repeating the process for each of the N nouns, we obtain their representations $\{\boldsymbol{\tau}_n\}_{n=1}^{N}$. The loss associated with the self-attention strategy is defined by

$$\mathcal{L}_e = \frac{1}{N} \sum_{n=1}^{N} -\log\left(\langle \mathbf{e}_{p_n}, W_e \boldsymbol{\tau}_n \rangle\right), \tag{3}$$

where W_e is the matrix of trainable parameters. $\mathcal{L}_e$ is the expression loss that aims to help each noun (entity) find other related words (descriptions) in the expression. It models the relationship of the n-th noun with other words in the expression through soft attention weights (*i.e.*, $\alpha_n = \{\alpha_{n,\ell}\}_{\ell=1}^{L}$). In other words, in the weight group of the n-th noun, α_n, the higher the correlation of a word to the n-th noun is, the larger weight this word is assigned.

C. Graph Collaborative Attention Module

Once obtaining the expression representations, we proceed to the next stage of constructing the graphs encoding the relationships between the objects in the image and the entities in the expression for the following reasoning and matching procedures. For this purpose, we construct a pair of complete graphs, the appearance graph $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$ and the categorical graph $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$, for appearance and semantic encoding, respectively. The appearance graph $\mathcal{G}_a$ is built by referring to the appearance of the object candidates and the expression. The categorical graph $\mathcal{G}_c$ is constructed similarly but replaces the appearance of the object candidates with their categorical labels, where the label of each proposal is obtained by leveraging the zero-shot, image-to-text matching capability of the CLIP model.

Specifically, suppose that we are given an image I with K detected object candidates as well as a predefined label set consisting of all possible categorical labels, including the base and unseen classes. The image and text encoders of the CLIP model are employed to extract the features of each proposal in the image and the features of each categorical label in the label set, respectively. It follows that cross-modal matching can be carried out for each proposal by finding its corresponding target label with the highest cosine similarity to its appearance features. Like the graph-based REC method [28], our method also maintains an appearance graph $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$ and a categorical graph $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$ for representing the visual appearance and categorical labels of the K object candidates, respectively. Nevertheless, the node features, $\mathcal{V}_a = \{v_i^a\}_{i=1}^{K}$ and $\mathcal{V}_c = \{v_i^c\}_{i=1}^{K}$, and the edge features, $\mathcal{E}_a = \{e_{ij}^a\}_{i,j=1}^{K}$ and $\mathcal{E}_c = \{e_{ij}^c\}_{i,j=1}^{K}$, are developed based on CLIP for carrying out zero-shot REC. They are elaborated as follows.

a) Node Features and Node Weights: Both the appearance graph $\mathcal{G}_a$ and the categorical graph $\mathcal{G}_c$ have K nodes, where K is the number of object candidates. A node in $\mathcal{G}_a$ encodes the appearance of its corresponding proposal and its matched noun in the expression, while a node in $\mathcal{G}_c$ considers the categorical semantics of the proposal as well as the matched noun. We first describe the node features of the appearance graph $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$. For the i-th node in $\mathcal{G}_a$, its features $v_i^a \in \mathcal{V}_a$ are computed via multilayer perceptron (MLP):

$$v_i^a = \mathbf{MLP}_a\left(\left[\mathbf{o}_i; \mathbf{b}_i; \boldsymbol{\tau}_{\pi_i^a}\right]\right) + \epsilon_a, \tag{4}$$

where multilayer perceptron $\mathbf{MLP}_a(\cdot)$ is composed of the fully-connected layers, and ϵ_a is a trainable bias. They are derived to adapt the concatenated feature vector $[\mathbf{o}_i; \mathbf{b}_i; \boldsymbol{\tau}_{\pi_i^a}]$ for reasoning and matching.

In the input to $\mathbf{MLP}_a(\cdot)$, the first component $\mathbf{o}_i$ is the visual features of the i-th object candidate and is extracted by the CLIP visual encoder. The second component $\mathbf{b}_i = W_b\left[x_i, y_i, w_i, h_i, w_i h_i\right]$ contains the box features of the i-th proposal, where W_b is a matrix of trainable parameters, (x_i, y_i) is the normalized coordinates of the proposal center, and w_i, h_i, and $w_i h_i$ are the normalized width, height, and area, respectively. The third component $\boldsymbol{\tau}_{\pi_i^a}$ contains the features of the π_i^a-th noun in the expression, where the noun features are calculated via (2), and among the N nouns in the expression, the π_i^a-th noun best matches the i-th proposal. The following describes how to find the best-match noun for the i-th proposal.

For the i-th proposal in the image and the n-th noun in the expression, their degree of matching is defined by

$$s_{i,n}^a = W_{a,2}\left[\tanh\left(W_{a,1}\left[\mathbf{o}_i; \mathbf{b}_i\right] + W_\tau \boldsymbol{\tau}_n\right)\right], \tag{5}$$

where $W_{a,1}$, $W_{a,2}$, and W_τ are three matrices of trainable parameters. It follows that the best-match noun $\boldsymbol{\tau}_{\pi_i^a}$ of the

i-th proposal is identified by

$$\pi_i^a = \underset{n}{\operatorname{argmax}} \{s_{i,n}^a | 1 \leq n \leq N\}, \tag{6}$$

where N denotes the number of nouns in the expression.

The node features of the appearance graph, $\mathcal{V}_a = \{v_i^a\}_{i=1}^K$, can be computed by repeating the procedure of (4) for each proposal i. To emphasize the nodes that better match nouns in the expression, we associate each node v_i^a with an additional weight w_i^a. Specifically, the node weight is defined by

$$w_i^a = \frac{\exp(s_{i,\pi_i^a}^a)}{\sum_{k=1}^K \exp(s_{k,\pi_k^a}^a)}. \tag{7}$$

In (4), (6), and (7), we have respectively computed the node features v_i^a, the index of the best-matched noun π_i^a, and the weight w_i^a for each node i in the appearance graph $\mathcal{G}_a$. For each node i in the categorical graph $\mathcal{G}_c$, its node features v_i^c, index of the best-match noun π_i^c, and weight w_i^c can be computed in the same way, except the visual features $\mathbf{o}_i$ which are replaced by the textual features $\boldsymbol{\zeta}_i$ obtained by applying the CLIP textual encoder to the i-th object candidate.

b) Edge Features and Edge Weights: Both $\mathcal{G}_a$ and $\mathcal{G}_c$ are fully-connected graphs initially. We first describe the edge features $e_{ij}^a \in \mathcal{E}_a$ in the appearance graph $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$. Like [29], our method considers all features of the two nodes that e_{ij}^a connects when deriving its features, namely

$$e_{ij}^a = W_e^a[\mathbf{o}_i; \mathbf{o}_j; \mathbf{b}_i; \mathbf{b}_j; \boldsymbol{\tau}_{\pi_i^a}; \boldsymbol{\tau}_{\pi_j^a}], \tag{8}$$

where W_e^a is a matrix of trainable parameters.

Edges that are relevant to the expression need to be emphasized. The degree of relevance between edge e_{ij}^a and the expression $\mathbf{q}$ is estimated via

$$\gamma_{ij}^a = W_q^a[\tanh(W_q\mathbf{q} + W_e^a e_{ij}^a)], \tag{9}$$

where W_q^a, W_q and W_e^a are the matrices of trainable parameters. When aggregating features from the i-th node, the weights of the edges linking this node are normalized and set to

$$w_{ij}^a = \frac{\exp(\gamma_{ij}^a)}{\sum_{k=1, i\neq k}^K \exp(\gamma_{ik}^a)} \cdot 1_{[i\neq j]}, \tag{10}$$

where $1_{[\cdot]}$ is the indicator function.

Previous methods [28], [29] show that in two-stage methods, the output of the first-stage procedure includes many irrelevant object candidates that introduce a lot of noise. This unfavorable effect can be alleviated by setting the edge weight w_{ij}^a to 0 if its original value is less than a pre-defined threshold. In this work, we empirically set the threshold to $1/K$.

For the appearance graph $\mathcal{G}_a$, we have described the computation of the edge features e_{ij}^a in (8) and the edge weight w_{ij}^a in (10). For the categorical graph $\mathcal{G}_c$, the features e_{ij}^c and the weight w_{ij}^c of each edge can be computed in the same way, except for all involved visual features which are replaced by the corresponding textual features extracted by using the CLIP textual encoder.

D. Multi-step Reasoning Module

There exist contextual relationships among the entities of a complex expression. Those relationships can be leveraged to more precisely locate the target object described by the expression. With the aid of graph-based structures, those relationships can be effectively explored through multiple message-passing steps among the nodes of the graph.

Specifically, once the graphs are constructed, we perform multi-step reasoning by using a T-layer GCN (graph convolutional network) where T corresponds to the number of the reasoning steps on $\mathcal{G}_a$ and $\mathcal{G}_c$ for information aggregation. In the following, We first describe the reasoning details on the appearance graph $\mathcal{G}_a$. The node features, node weights, edge features, and edge weights on $\mathcal{G}_a$ at the t-th reasoning step are denoted by $\{v_i^a(t)\}_{i=1}^K$, $\{w_i^a(t)\}_{i=1}^K$, $\{e_{ij}^a(t)\}_{i,j=1}^K$, and $\{w_{ij}^a(t)\}_{i=1}^K$, respectively. In the beginning *i.e.*, $t = 0$, we initialize them to what we compute in (4), (7), (8), and (10), respectively. That is, $v_i^a(0) = v_i^a$, $w_i^a(0) = w_i^a$, $e_{ij}^a(0) = e_{ij}^a$, and $w_{ij}^a(0) = w_{ij}^a$.

We adopt a similar message passing strategy to [29] on $\mathcal{G}_a$, where the i-th node feature of $\mathcal{G}_a$ at the t-th reasoning step is calculated by

$$v_i^a(t) = W_t\left(\widetilde{v}_i^a(t) + \widehat{v}_i^a(t) + \boldsymbol{\theta}_t\right) + v_i^a(t-1), \tag{11}$$

where W_t is a matrix of trainable parameters and $\boldsymbol{\theta}_t$ is a vector of trainable parameters, and $\widetilde{v}_i^a(t)$ and $\widehat{v}_i^a(t)$ are the aggregated node features from the connected neighboring nodes and the transformed features for the self-loop of the i-th node at the $(t-1)$-th reasoning step, respectively. More Specifically,

$$\begin{aligned}
\widetilde{v}_i^a(t) &= \sum_{j=1, j\neq i}^K w_{ij}^a(t-1)\left(\widetilde{W}_t w_j^a(t-1) v_j^a(t-1) + \widetilde{\boldsymbol{\theta}}_t\right), \\
\widehat{v}_i^a(t) &= \widehat{W}_t w_i^a(t-1) v_i^a(t-1) + \widehat{\boldsymbol{\theta}}_t,
\end{aligned} \tag{12}$$

where $\widetilde{W}_t$ and $\widehat{W}_t$ are two matrices of trainable parameters, and $\widetilde{\boldsymbol{\theta}}_t$ and $\widehat{\boldsymbol{\theta}}_t$ are vectors of trainable parameters.

After a forward pass of a layer of graph convolutions, each compound node of $\mathcal{G}_a$ aggregates the information from other nodes where a compound node that is more relevant to the expression should be assigned a larger weight. Accordingly, the node and edge weights of $\mathcal{G}_a$ (*i.e.*, $w_i^a(t)$ and $w_{ij}^a(t)$) are updated as follows to take the expression representation $\mathbf{q}$ into consideration as well:

$$\begin{aligned}
s_i^a(t) &= W_{a,2}^{(t)}\left[\tanh\left(W_q\mathbf{q} + W_{a,1}^{(t)} v_i^a(t)\right)\right], \\
\gamma_{ij}^a(t) &= W_{e,2}^{(t)}\left[\tanh\left(W_q\mathbf{q} + W_{e,1}^{(t)}\left[v_i^a(t); v_j^a(t); e_{ij}^a\right]\right)\right],
\end{aligned} \tag{13}$$

where $W_{a,2}^{(t)}$, $W_{a,1}^{(t)}$, $W_{e,1}^{(t)}$, and $W_{e,2}^{(t)}$ are trainable matrices of parameters. $s_i^a(t)$ and $\gamma_{ij}^a(t)$ are the relevance scores of nodes and edges of $\mathcal{G}_a$ at the t-th reasoning step, respectively. When performing the feature aggregation of the i-th node at the $(t+1)$-th step, the node weight and the corresponding edge weights of the i-th node are further normalized by the scores of all nodes, and all edges connecting the i-th node, respectively, *i.e.*, similar to (7) and (10).

The aforementioned process is repeated T times for the T-step reasoning. The node features, node weights, edge features,

and edge weights of the categorical graph $\mathcal{G}_c$ can be obtained in the same way as $\mathcal{G}_a$, except the visual features which are replaced by the corresponding textual features extracted by using the CLIP textual encoder.

E. Loss Function and Matching Module

After the T-step reasoning, we are able to compute the matching scores between the expression and the compound nodes of both graphs, $\mathcal{G}_a$ and $\mathcal{G}_c$, for identifying the target object specified by the expression. In graph-based representations, those nodes with higher matching scores are more relevant to the expression. In this work, the matching scores of the i-th compound node in graph $\mathcal{G}_*$ are calculated by

$$S_i^* = \left\langle \frac{W_c^* v_i^*(T)}{\|W_c^* v_i^*(T)\|}, \frac{W_d^* \mathbf{q}}{\|W_d^* \mathbf{q}\|} \right\rangle, \tag{14}$$

where $* \in \{a, c\}$, $v_i^*(T)$ is the features of the i-th node at the T-th iteration, and W_c^* and W_d^* are two trainable matrices of parameters. $||\cdot||$ means the L_2-norm of a vector.

The scores of the ground-truth (GT) nodes in $\mathcal{G}_a$ and $\mathcal{G}_c$ are denoted by S_{GT}^a and S_{GT}^c, and they should be maximized during the training phase. To jointly consider two attention graphs, we define the loss function of this work as follows:

$$\mathcal{L} = -\log\left(P_{\mathrm{GT}}\right) \tag{15}$$

where

$$P_{\mathrm{GT}} = \frac{\exp\left(S_{\mathrm{GT}}^a + S_{\mathrm{GT}}^c\right)}{\sum_{k=1}^{K} \exp\left(S_k^a + S_k^c\right)}. \tag{16}$$

a) Network Optimization: During training, we first use $\mathcal{L}_{\mathrm{e}}$ in (3) to train our proposed language parser. Then, we fix the trained parameters of the language parser and learn the rest modules in CLIPREC via $\mathcal{L}$ in (15).

b) Matching: During inference, for each test sample, we construct the two attention graphs by the same procedures as described above. The correspondence between the unseen object categories and candidate categorical labels can still be matched by using the features from the image and text encoders of the pre-trained CLIP model. The two graphs are fed into the multi-step reasoning module for information aggregation. A matching score is obtained for each compound node of both graphs by computing its similarity to the expression via (14). The object corresponding to the node with the highest score is retrieved as the target, and the task of (zero-shot) REC is completed.

IV. Experiments

We first focus on zero-shot REC and evaluate our method on the Visual Genome [39] and Flickr30K [40] datasets by following the setup of [10] and on our collected RefCOCOZ and RefCOCOZ+ datasets. Then, we conduct ablation studies and show some qualitative results. Finally, we show the evaluation results of the proposed CLIPREC on RefCOCO [41], RefCOCO+ [41], and RefCOCOg [42] for the standard REC task. Due to the limit of the space, we only describe evaluation results of the zero-shot REC task. The descriptions of the standard REC task are provided in the supplementary results.

A. Datasets and Implementation Details

We introduce the datasets used for standard and zero-shot REC and describe the implementation details of our method.

a) Datasets for standard REC: **RefCOCO:** The RefCOCO dataset contains 142,210 expressions for 50,000 objects in 19,994 images where these expressions are collected via an interactive game interface. In [41], this dataset is split into four parts: 120,624, 10,834, 5,657, and 5,095 expression-referent pairs for training, validation, testA, and testB, respectively. TestA includes images of multiple people, while testB covers images of multiple other objects.

RefCOCO+: The RefCOCO+ dataset has 141,564 expressions for 49,856 objects in 19,992 images collected via an interactive game interface. Different from those in RefCOCO, the expressions in RefCOCO+ do not contain any descriptions of absolute locations. In [41], RefCOCO+ is also split into four parts: 120,191, 10,758, 5,726, and 4,889 expression-referent pairs for training, validation, testA and testB, respectively.

RefCOCOg: The RefCOCOg dataset contains 95,010 long expressions for 49,822 objects in 25,799 images collected in a non-interactive setting. In this dataset, 80,512 expression-referent pairs are used for training, 4,896 pairs for validation, and 9,602 pairs for testing. The RefCOCO, RefCOCO+, and RefCOCOg datasets [42] are collected from the MSCOCO dataset [44] of 80 object categories.

b) Datasets for zero-shot REC: The datasets used for zero-shot REC are collected from Flickr30K [40] and Visual Genome [45], RefCOCO and RefCOCO+.

Flickr30K: It contains 31,783 images, where each image is described in 5 sentences. For each sentence, the corresponding referred objects are from one of 7 common object categories and another category that does not belong to the 7 object categories.

Visual Genome: It consists of 108,077 images, 5.4 million region descriptions, 1.7 million visual question answers, 3.8 million object instances in 33,877 categories, 2.8 million attributes, and 2.3 million relationships.

We follow the protocol in [10] for zero-shot REC where we split Flickr30K into the Flickr-Split-0 and Flickr-Split-1 subsets and split Visual Genome into the VG-Split-2 and VG-Split-3 subsets to cover the following four cases of zero-shot REC. (1) Case 0: In Flickr-Split-0, the referred noun in its test set is not included in its training set; (2) Case 1: In Flickr-Split-1, the categories of the referred objects in its test set are not covered by its training set; (3) Case 2: In VG-Split-2, the objects semantically close to the referred objects in the test set only appear in the training set; (4) Case 3: In VG-Split-3, the objects semantically close to the referred objects in the test set appear not only in the training set but also in the test set. More details about the four cases of zero-shot REC can be found in [10].

In addition, we also create the RefCOCOZ and RefCOCOZ+ datasets by collecting images from RefCOCO [41] and RefCOCO+ [41] respectively where the images of RefCOCO and RefCOCO+ are collected from MSCOCO dataset [44] of 80 object categories. The setting of zero-shot REC on the two datasets is the combination of case 0 and case 1 [10], hence termed case 01. For the two datasets, we choose 51

TABLE I
COMPARISON OF OUR COLLECTED REFCOCOZ(+) AND REFCOCO(+). THE TARGET OBJECTS OF THE TEST SET OF REFCOCOZ(+) ARE ALSO FROM FOUR CATEGORIES: 16,882 (16,573), 8,556 (8,243), 3,988 (3,560), AND 6,636 (6,058) FOR PERSON, ANIMAL, VEHICLE, AND FOOD, RESPECTIVELY. IN ADDITION, THE TARGET OBJECTS (*i.e.*, THE NOUNS IN THE EXPRESSION AND CATEGORIES OF OBJECTS IN AN IMAGE) IN THE TEST SET OF REFCOCOZ(+) ARE NOT PRESENT IN THE TRAIN AND VALIDATION SETS OF REFCOCOZ(+).

RefCOCOZ			RefCOCO			RefCOCOZ+				RefCOCO+			
train	val	test	train	val	test	train	val	testA	testB	train	val	testA	testB
15,137	2,083	35,862	120,624	10,834	5,657	5,095	13,858	1,667	29,034	120,191	10,758	5,726	4,889

TABLE II
PERFORMANCE OF ZERO-SHOT REC USING THE PROPOSED CLIPREC AND FIVE COMPETING METHODS ON FLICKR-0, FLICKR-1, VG-SPLIT-2, AND VG-SPLIT-3, WHERE "B" AND "UB" DENOTE THE BALANCED AND UNBALANCED SETS, RESPECTIVELY, AND "0.3" AND "0.5" MEAN THE CRITERIA IOU$\geq$ 0.3 AND IOU$\geq$ 0.5, RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	Back-bone	Flickr-Split-0	Flickr-Split-1	VG-2B		VG-2UB		VG-3B		VG-3UB	
				0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
QRG [43]	VGG	35.62	24.42	13.17	7.64	12.39	7.15	14.21	8.35	13.03	7.52
ZSGNet [10]	VGG	39.32	29.35	17.09	11.02	16.48	10.55	17.63	11.42	17.35	10.97
	R50	43.02	31.23	19.95	12.90	19.12	12.37	20.77	13.77	19.72	12.82
CLIP-DGA	R101	56.53	50.67	63.08	60.21	65.57	63.38	64.14	60.63	63.25	61.19
CLIP-CMRIN	R101	56.07	51.25	64.35	60.05	65.48	63.96	64.35	61.08	64.75	62.56
CLIP-SGMN	R101	56.87	50.06	63.07	61.14	65.35	62.58	63.97	61.13	64.54	60.67
CLIPREC w/o MLP	R50	58.56	51.71	65.26	64.82	68.05	66.15	68.54	64.98	66.82	65.07
CLIPREC	R50	64.85	55.56	72.63	69.56	73.74	70.10	74.05	71.86	72.96	70.33
CLIPREC w/o MLP	R101	59.23	53.12	66.14	65.50	69.37	67.68	68.89	65.52	67.74	66.83
CLIPREC	R101	**66.02**	**57.24**	**73.65**	**71.16**	**75.67**	**72.22**	**76.85**	**73.08**	**73.98**	**72.32**

out of 80 classes as the seen classes and re-organize the rest into four subsets for the unseen, including person (1 class), animal (10 classes), vehicle (8 classes), and food (10 classes) due to their majorities in the two datasets. We utilize the natural language toolkit (NLTK) [46] to analyze the expression and facilitate the curation process. Note that each sample in RefCOCO and RefCOCO+ contains the object candidates, categories, and an expression. For the test set, we select the samples whose target objects are related to person, animal, vehicle, or food classes to form RefCOCOZ-test and RefCOCOZ+-test. The numbers of test samples of the person, animal, vehicle, and food subsets in RefCOCOZ-test are 16,882, 8,556, 3,988, and 6,636, respectively. The corresponding four numbers in RefCOCOz+-test are 16,573, 8,243, 3,560, and 6,058, respectively. We prepare the training data by removing all the samples whose expressions contain the nouns related to subsets person, animal, vehicle, or food in RefCOCO and RefCOCO+. For the rest samples, if their object candidates and corresponding categories are related to subsets person, animal, vehicle, or food, we also remove them. The remaining samples form RefCOCOZ-train and RefCOCOZ+-train, and the numbers of samples in RefCOCOZ-train and RefCOCOZ+-train are 15,137 and 13,858, respectively. The comparision between RefCOCOZ(+) and RefCOCO(+) are shown in Table I.

c) Implementation details: Our CLIPREC [1] is a two-stage method. For the REC task shown in [28], there are two setups for training: the ground-truth object strategy and the detected object strategy. Both ground-truth and detected objects are used to construct the graphs in the ground-truth object strategy. By contrast, in the detected object strategy, the graphs are constructed only with the detected objects. The detected objects with the highest IoU (intersection over union) values between the detected object and the ground-truth are considered the ground-truth objects. During prediction in the ground-truth object strategy, the ground-truth object with the highest score is considered correct. In the detected object strategy, if the IoU value between the object with the highest score and the ground-truth object is larger than 0.5, the prediction is considered correct.

For zero-shot REC, we train the region proposal network by following the same training settings as [10] on Flickr30k and Visual Genome. For RefCOCOZ and RefCOCOZ+, we use MSCOCO to train the region proposal network to obtain the detected objects. It is worth noting that the detected objects labeled as "background" are removed.

In the setting of zero-shot evaluation, because the target objects are from unseen classes, we employ the detected object strategy for training. In addition, during inference, we use the image (*i.e.*, ResNet50 and ResNet101) and text encoders of the pre-trained CLIP model to extract the embedded visual and textual features from the input image and expression as well as match the object candidates with the corresponding category labels. The manual prompt of the text encoder is utilized from the zero-shot module of CLIP (i.e., "a photo of a [category]"). Following [10], we use the top-1 accuracy (%) as the evaluation metric. The number of epochs, the batch size, and the learning rate for training are set to 30, 32, and 10^{-4}, respectively. In addition, the number of reasoning steps, *i.e.*, the number T of layers in GCN, is set to 3.

[1]The source code will be released upon acceptance.

TABLE III
PERFORMANCE OF ZERO-SHOT REC WITH CLIPREC AND FOUR COMPETING METHODS ON REFCOCOZ AND REFCOCOZ+, WHERE THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Methods	Back-bone	RefCOCOZ					RefCOCOZ+				
		average	person	animal	vehicle	food	average	person	animal	vehicle	food
ZSGNet [10]	R50	42.59	45.53	34.22	49.98	40.65	29.27	30.50	24.47	33.41	28.70
RECLIP [24]	R50	45.77	47.75	45.34	41.83	48.16	44.81	44.65	42.59	42.37	49.63
CLIP-DGA	R101	67.77	66.84	63.13	68.75	72.39	48.68	53.56	39.15	47.80	54.23
CLIP-CMRIN	R101	67.69	65.25	63.53	69.13	72.88	48.47	52.46	39.83	47.56	55.06
CLIP-SGMN	R101	68.02	65.95	64.12	68.98	73.03	48.62	53.12	37.93	48.29	55.15
CLIPREC w/o **MLP**	R50	73.08	71.75	68.20	75.05	77.34	54.02	58.23	44.17	53.68	60.03
CLIPREC	R50	78.80	78.07	72.74	81.39	82.98	58.64	64.50	46.48	57.92	65.68
CLIPREC w/o **MLP**	R101	74.16	72.28	69.64	75.82	78.92	55.24	59.54	45.59	54.48	61.26
CLIPREC	R101	**79.57**	**78.24**	**73.69**	**82.56**	**83.77**	**60.47**	**64.53**	**48.45**	**60.92**	**67.97**

TABLE IV
ABLATION STUDY RESULTS WITH $T = 3$ WITH DIFFERENT LOSS TERMS FOR ZERO-SHOT REC

Loss	Flickr-Split-0	Flickr-Split-1	VG-2B		VG-2UB		VG-3B		VG-3UB		RefCOCOZ	RefCOCOZ+
			0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5	avg	avg
$\mathcal{L}_a$	64.35	55.06	71.80	68.74	73.60	70.15	73.51	71.14	71.64	70.08	78.03	58.52
$\mathcal{L}_c$	59.65	50.23	66.24	65.89	68.24	65.70	68.82	67.29	68.03	66.62	71.20	53.63
$\mathcal{L}$	**66.02**	**57.24**	**73.65**	**71.16**	**75.67**	**72.22**	**76.85**	**73.08**	**73.98**	**72.32**	**79.57**	**60.47**

TABLE V
ABLATION STUDY RESULTS WITH $T = 3$ WITH AND WITHOUT $\mathcal{L}_e$ AND TEXTUAL FEATURE OF EXPRESSION IN THE NODES OF CONSTRUCTED GRAPHS RESPECTIVELY FOR ZERO-SHOT REC

$\mathcal{L}_e$	Text	Flickr-Split-0	Flickr-Split-1	VG-2B		VG-2UB		VG-3B		VG-3UB		RefCOCOZ	RefCOCOZ+
				0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5	avg	avg
	✓	65.58	56.67	72.86	70.07	73.96	71.18	76.05	72.26	73.16	71.79	79.28	59.94
✓		63.93	55.28	70.18	69.67	73.46	70.17	74.83	71.96	70.75	69.04	77.75	58.20
✓	✓	**66.02**	**57.24**	**73.65**	**71.16**	**75.67**	**72.22**	**76.85**	**73.08**	**73.98**	**72.32**	**79.57**	**60.47**

TABLE VI
ABLATION STUDY RESULTS WITH DIFFERENT NUMBERS OF GCN LAYERS FOR ZERO-SHOT REC

No. layers (T)	Flickr-Split-0	Flickr-Split-1	VG-2B		VG-2UB		VG-3B		VG-3UB		RefCOCOZ	RefCOCOZ+
			0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5	avg	avg
0	62.24	50.56	69.21	67.63	68.01	66.52	72.33	70.18	68.63	67.24	75.80	57.18
1	65.16	54.74	71.52	69.65	73.02	70.86	74.46	72.36	72.15	70.21	78.83	60.04
2	**66.27**	56.67	73.04	71.12	**75.83**	71.84	76.63	**73.28**	73.93	71.16	**79.62**	59.93
3	66.02	**57.24**	**73.65**	**71.16**	75.67	**72.22**	**76.85**	73.08	**73.98**	**72.32**	79.57	**60.47**
4	64.58	56.60	72.24	69.78	74.05	70.93	75.56	71.96	72.25	71.64	78.98	60.21

TABLE VII
EXPERIMENTAL RESULTS IN THE ZERO-SHOT SETTING USING DIFFERENT MANUAL PROMPTS. P1 AND P2 REPRESENT THE PROMPT "THIS IS A PHOTO OF [CATEGORY]" AND "[CATEGORY]", RESPECTIVELY.

Methods	Backbone	RefCOCOZ avg	RefCOCOZ+ avg
CLIPREC with P1	R101	79.37	60.12
CLIPREC with P2	R101	**80.04**	60.33
CLIPREC	R101	79.57	**60.47**

TABLE VIII
EXPERIMENTAL RESULTS IN THE ZERO-SHOT SETTING USING THE UNSUPERVISED CLIPREC AND CLIPREC W/O CLIP FEATURES

Methods	Backbone	Flickr-Split-0	Flickr-Split-1
CLIPREC w/o CLIP features	R101	51.23	47.66
CLIPREC	R101	**66.02**	**57.24**

B. Comparisons with State-of-the-Arts for Zero-Shot REC

We evaluate the generalization ability of CLIPREC for zero-shot REC where the expressions contain the instances of unseen object categories during the inference phase. We apply CLIPREC on Flickr-Split-0, Flickr-Split-1, VG-Split-2, and VG-Split-3 using the same evaluation protocols as that in [10]. To measure the effect of using CLIP for feature extraction, we create three new baselines, called CLIP-DGA, CLIP-SGMN, and CLIP-CMRIN, by replacing the original features with the visual and textual features extracted by CLIP in three graph-based REC methods, DGA [28], SGMN [47], and CMRIN [29], respectively. We create a baseline CLIPREC w/o **MLP** for our method. In this baseline, the features extracted from the CLIP model are used directly without using MLP for adaptation introduced in Sec. III-C. ResNet50 and ResNet101 are adopted as the backbones in all competing methods for a fair comparison.

The evaluation results are shown in Table II, from which we find that the performances of CLIPREC w/o $\mathbf{MLP}$ with backbone ResNet101 are better than CLIP-DGA, CLIP-SGMN, and CLIP-CMRIN. It shows that CLIP in our method is well-designed for REC and gives promising results with the same CLIP features. More importantly, CLIPREC achieves a significant performance gain over CLIPREC w/o $\mathbf{MLP}$. It indicates that our method effectively adapts the CLIP features to the REC task. Also, CLIPREC outperforms the previous state-of-the-art method, ZSGNet-ResNet50, for zero-shot REC.

We further evaluate CLIPREC on the collected RefCOCOZ and RefCOCOZ+ datasets for case 01 zero-shot REC [10]. In compared methods, RECLIP is the latest released zero-shot REC method that it locates the object described by expression by directly feeding the object candidates, noun-phrases, and spatial-relation words of expression into the image encoder of CLIP, the text encoder of CLIP, and a spatial-relation parser, respectively. More importantly, RECLIP is only suitable for expressions that include spatial-relation words such as "left," "right," "top," and "down". The results report in Table III show that both CLIPREC and CLIPREC w/o $\mathbf{MLP}$ considerably outperform ZSGNet and RECLIP. Specifically, in RefCOCOZ+ where spatial-relation words are not included in the expressions, CLIPREC w/o MLP using R50 performs on average 9.21% better than RECLIP. These results demonstrate that our method can handle more complex sentence structures. Meanwhile, the average top-1 accuracy of CLIPREC w/o $\mathbf{MLP}$ with backbone ResNet101 is approximately 5% higher than CLIP-DGA, CLIP-SGMN, and CLIP-CMRIN. The results demonstrate that the CLIP models can well serve the purpose of feature extraction for image-text matching and do not perform any reasoning and understanding of the structure and semantics of the given image and expression. Finally, we show that not only rich information encoded in the CLIP features but also feature adaptation enabled by multilayer perceptron realize more accurate zero-shot REC. When $\mathbf{MLP}$ is appended to the image and text encoders of CLIP, the performance gains of CLIPREC over its variant CLIPREC w/o $\mathbf{MLP}$ are on average about 7%, which validates that our design of $\mathbf{MLP}$ for feature adaptation is effective and can make the most of CLIP for zero-shot REC.

C. Ablation Studies

To validate the effectiveness of the appearance and categorical graphs in CLIPREC, we evaluate and compare the results of using three different losses, $\mathcal{L}_a$, $\mathcal{L}_c$, and $\mathcal{L}$, for performance verification. Note that using $\mathcal{L}_a$ (or $\mathcal{L}_c$) only stands for only considering the visual (or categorical) graph construction module while using $\mathcal{L}$ stands for considering both $\mathcal{L}_a$ and $\mathcal{L}_c$. The loss $\mathcal{L}$ in (15) that CLIPREC adopts is a function of the probability P_{GT} in (16), where the probability P_{GT} is estimated based on the scores computed in both the appearance graph and the categorical graph. The losses $\mathcal{L}_a$ and $\mathcal{L}_c$ are similar to $\mathcal{L}$, but they consider just the scores computed in the appearance graph and the categorical graph, respectively. We adopt CLIPREC with the ResNet101 backbone as the baseline for this ablation study. The results reported in Table IV show that, although the performance of CLIPREC with $\mathcal{L}_a$ is better than that with $\mathcal{L}_c$, using both graphs with $\mathcal{L}$ can still boost the performance. This confirms that the categorical graph encodes information complementary to the appearance graph, leading to further performance improvement.

In the case of using $\mathcal{L}$, we evaluate CLIPREC with and without loss $\mathcal{L}_a$ and textual features of expression in nodes of graphs, respectively. The results reported in Table V show that the performance of CLIPREC drops sightly if $\mathcal{L}_e$ is removed. However, if the textual features of expression are removed, the performance of CLIPREC drops by about 3%. The results demonstrate that using fusing the textual features of expression to the nodes of graphs and adding $\mathcal{L}_e$ help improve performances greatly.

Next, we evaluate the proposed CLIPREC with different numbers of GCN layers, *i.e.*, the numbers of reasoning steps, and report the results in Table VI. After graph initialization (*i.e.*, $T = 0$), the performance of CLIPREC serves as the baseline where the nodes of two attention graphs in CLIPREC only contain the initial node features without exploiting any relationship between objects. Through feature aggregation via GCN (*i.e.*, $T > 0$), the best performance of CLIPREC is achieved when two or three layers of GCN are adopted. However, once using more layers starts to result in degraded performances. We conjecture the reason could be to aggregate the information from irrelevant objects.

We further evaluate the capability of CLIPREC for zero-shot REC on different manual prompts of the text encoder. Table VII shows that the performance of CLIPREC remains consistent regardless of whether prompt "This is a photo of [category]" or "[category]" is used, with performance levels similar to those achieved using the zero-shot module of CLIP.

Finally, we evaluate the capability of CLIPREC for zero-shot REC by keeping the language parser and replacing the visual and categorical features of CLIPREC with the visual features from faster R-CNN and GloVe features, respectively. Table VIII shows that the performance of CLIPREC decreases evidently because the domains of faster R-CNN, GloVe, and the language parser are less related. It is hard to project these three kinds of domains into a common domain.

D. Qualitative Results of Zero-Shot REC

In this subsection, we show some qualitative results of the proposed approach for zero-shot REC. The visualization samples in the five above-mentioned cases are shown in the first, second, and third rows of Fig. 3. In case 0, we find that CLIPREC can locate the correct object even though CLIPREC never learns the target nouns "ball" and "cattle." In case 1, although CLIPREC did not see the objects in the categories "people" and "vehicle", it can still locate the correct object. In case 2, we observe that CLIPREC can effectively locate the target objects even though the size of the target object is small. In case 3, it can be found that CLIPREC can locate the target object even when multiple objects and the target are from the same category in the image. We get similar results on our curated RefCOCOZ and RefCOCOZ+ that CLIPREC can locate the objects described by the expressions for the unseen

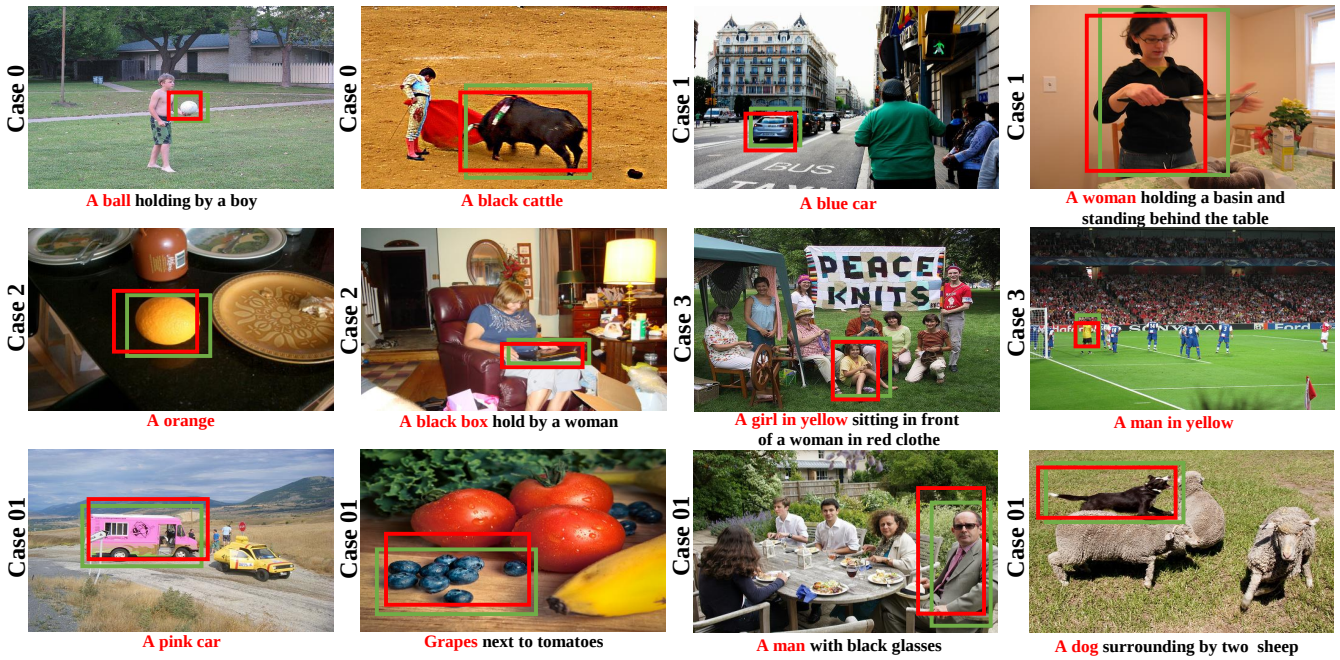


Fig. 3. Visualization results of CLIPREC in five zero-shot REC cases. The first two rows show the results in cases 0, 1, 2, and 3, respectively, while the third row shows the results in case 01. The bounding box in green and red represent the ground-truth bounding box and predicted bounding box, respectively.

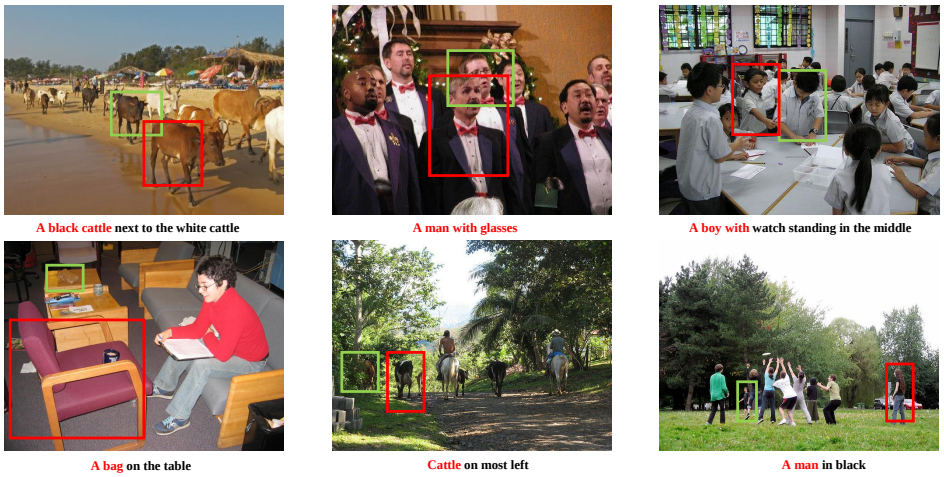


Fig. 4. Some failure results of CLIPREC. The bounding box in green and red represents the ground-truth bounding box and predicted bounding box, respectively.

objects of categories “people,” “animal,” “food,” and “vehicle” before. These results demonstrate the strength of CLIPREC for zero-shot REC.

In addition, Fig. 4 also illustrates some failure results. In the first row, for the case of ”A black cattle next to the white cattle,” the target object is surrounded by similar objects. Similarly, in the example of ”A man with glasses,” the ground-truth bounding box of the target person includes another person, and the watch in ”A boy with watch standing in the middle” is too small. These object relationship ambiguities caused by distractors of similar objects and mis-detections by the object detector can lead CLIPREC to make incorrect predictions. In the second row of Fig. 4, we observe that the detector struggles to locate the target object due to its small size, partial occlusion, or similar coloration with the background, leading to incorrect predictions.

These results demonstrate not only the effectiveness of the proposed approach but also how it improves the reasoning capability beyond the original CLIP model.

V. Conclusion

In this work, we proposed the CLIPREC method for zero-shot REC. By carefully leveraging the joint image-text feature space of the pre-trained CLIP model along with the feature adaptation layers, our method is able to effectively exploit the correlation between two graphs with the expression containing seen or unseen object categories for precise predictions. Our proposed CLIPREC further includes learnable layers to adapt the contrastively learned CLIP model to the

REC task, resulting in significant performance improvement. Extensive experimental and ablation results on the Flickr30K, Visual Genome, and our curated RefCOCOZ and RefCOCOZ+ datasets demonstrate the superior performance of CLIPREC with respect to the state-of-the-arts for zero-shot REC.

REFERENCES

[1] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2015, pp. 3668–3678.

[2] Y. Wang, J. Deng, W. Zhou, and H. Li, "Weakly supervised temporal adjacent network for language grounding," *IEEE Trans. Multimedia*, vol. 24, pp. 3276–3286, July 2022.

[3] Y. Qiao, C. Deng, and Q. Wu, "Referring expression comprehension: A survey of methods and datasets," *IEEE Trans. Multimedia*, vol. 23, pp. 4426–4440, December 2021.

[4] X. Zhang, Z. Wu, and Y.-G. Jiang, "Sam: Modeling scene, object and action with semantics attention modules for video recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 313–322, January 2022.

[5] L. Chen, W. Ma, J. Xiao, H. Zhang, and S.-F. Chang, "Ref-NMS: Breaking proposal bottlenecks in two-stage referring expression grounding," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, May 2020, pp. 1036–1044.

[6] C. Jing, Y. Wu, M. Pei, Y. Hu, Y. Jia, and Q. Wu, "Visual-semantic graph matching for visual grounding," in *Proc. ACM Multimedia Conf.(MM)*, October 2020, p. 4041–4050.

[7] X. Chen, L. Ma, J. Chen, Z. Jie, W. Liu, and J. Luo, "Real-time referring expression comprehension by single-stage grounding network," in *CoRR*, 2018.

[8] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "TransVG: End-to-end visual grounding with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 1769–1779.

[9] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr - modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 1780–1790.

[10] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *Proc. IEEE/CVF Int. Conf. Comput. vis.. (ICCV)*, October 2019, pp. 4694–4703.

[11] J. Wang and L. Specia, "Phrase localization without paired training examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2019, pp. 4663–4672.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Machine Learning Research*, July 2021, pp. 8748–8763.

[13] A. Kazi, L. Cosmo, S.-A. Ahmadi, N. Navab, and M. Bronstein, "Differentiable graph module (dgm) for graph convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–13, April 2022.

[14] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 26, 2013, pp. 935–943.

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 26, 2013, pp. 2121–2129.

[16] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do CIFAR-10 classifiers generalize to cifar-10?" in *CoRR*, 2018.

[17] A. Li, A. Jabri, A. Joulin, and L. van der Maaten, "Learning visual n-grams from web data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2017, pp. 4193–4202.

[18] Y. Jin, W. Jiang, Y. Yang, and Y. Mu, "Zero-shot video event detection with high-order semantic concept discovery and matching," *IEEE Trans. Multimedia*, vol. 24, pp. 1896–1908, April 2022.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2009.

[20] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Represent.(ICLR)*, 2022.

[21] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[22] M. Li, R. Xu, S. Wang, L. Zhou, X. Lin, C. Zhu, M. Zeng, H. Ji, and S.-F. Chang, "Clip-event: Connecting text and images with event structures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 16 420–16 429.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, July 2022.

[24] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach, "ReCLIP: A strong zero-shot baseline for referring expression comprehension," in *The Assn. Comput. Linguistics (ACL)*, may 2022, pp. 5198–5215.

[25] C. Liang, W. Wang, T. Zhou, and Y. Yang, "Visual abductive reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 15 565–15 575.

[26] Y. Liu, B. Wan, X. Zhu, and X. He, "Learning cross-modal context graph for visual grounding," in *Proc. AAAI Conf. Artif. Intell.(AAAI)*, April 2020, pp. 11 645–11 652.

[27] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. v. d. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019, pp. 1960–1968.

[28] S. Yang, G. Li, and Y. Yu, "Dynamic graph attention for referring expression comprehension," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2019, pp. 4644–4653.

[29] ——, "Graph-structured referring expression reasoning in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020, pp. 9952–9961.

[30] C. Gao, J. Chen, S. Liu, L. Wang, Q. Zhang, and Q. Wu, "Room-and-object aware knowledge reasoning for remote embodied referring expression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2021, pp. 3064–3073.

[31] C. Liang, Y. Wu, Y. Luo, and Y. Yang, "Clawcranenet: Leveraging object-level relation for text-based video segmentation," in *CoRR*, 2021.

[32] C. Liang, W. Wang, T. Zhou, J. Miao, Y. Luo, and Y. Yang, "Local-global context aware transformer for language-guided video segmentation," in *CoRR*, 2022.

[33] Z. Li, L. Yao, X. Zhang, X. Wang, S. Kanhere, and H. Zhang, "Zero-shot object detection with textual descriptions," in *Proc. AAAI Conf. Artif. Intell.(AAAI)*, July 2019, pp. 8690–8697.

[34] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "Promptdet: Towards open-vocabulary detection using uncurated images," in *Proc. Eur. Conf. Comput. vis.. (ECCV)*, November 2022, pp. 701–717.

[35] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2330–2343, February 2018.

[36] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2018, pp. 1307–1315.

[37] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2017, pp. 804–813.

[38] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, July 2017, pp. 1115–1124.

[39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 06, p. 32–73, February 2017.

[40] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, December 2015, pp. 2641–2649.

[41] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Conf. Empir. Methods in Nat.Language Proc (EMNLP)*, October 2014, p. 787–798.

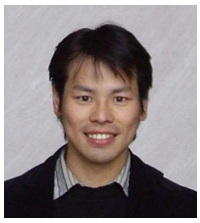
[42] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016, pp. 11–20.

[43] K. Chen, R. Kovvuri, and R. Nevatia, "Query-guided regression network with context policy for phrase grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2017, pp. 824–832.
[44] J. Pont-Tuset and L. Van Gool, "Boosting object proposals: From pascal to coco," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, February 2015, pp. 1546–1554.
[45] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 24, 2011, pp. 1143–1151.
[46] S. Bird and E. Loper, "NLTK: The natural language toolkit," in *Proc. Assn. for. Comput. Linguistics (ACL)*, July 2004, p. 214–217.
[47] S. Yang, G. Li, and Y. Yu, "Relationship-embedded representation learning for grounding referring expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2765–2779, February 2021.

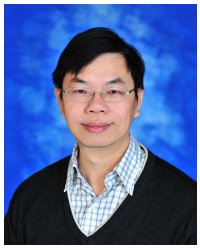
Jingcheng Ke received the B.S. degree in Computer Science and Engineering from Southwest Minzu University, Chengdu, China, in 2014, and the M.S. degree in Computer Science from Shaanxi Normal University, Xian, China, in 2018. He is currently working toward the ph.D. degree in Institute of Communications Engineering of National Tsing Hua University (NTHU), Hsinchu, Taiwan. His research interests include Referring Expression Comprehension, Diffusion Models, and Graph Neural Networks.

Jia Wang received the Ph.D. degrees from the Institute of Electronics, National Yang-Ming Chiao Tung University (NYCU), Taiwan in 2023, the M.S. degree in Aeronautics and Astronautics from the University of Electronic Science and Technology of China, Chengdu, China, in 2018, and the B.S. degree in Electrical Engineer from Southwest Minzu University, Chengdu, China, in 2014. Her research interests include machine learning, computer vision, and multimedia analytics.

Jun-Cheng Chen (Member, IEEE) is now an associate research fellow at the Research Center for Information Technology Innovation (CITI), Academia Sinica. He joined CITI as an assistant research fellow in 2019. He received the B.S. and M.S. degrees advised by Prof. Ja-Ling Wu in Computer Science and Information Engineering from National Taiwan University, Taiwan (R.O.C), in 2004 and 2006, respectively, where he received the Ph.D. degree advised by Prof. Rama Chellappa in Computer Science from University of Maryland, College Park, USA, in 2016. From 2017 to 2019, he was a postdoctoral research fellow at University of Maryland Institute for Advanced Computer Studies. His research interests include computer vision, machine learning, deep learning and their applications to biometrics, such as face recognition/facial analytics, activity recognition/detection in the visual surveillance domain, etc. His works have been recognized in prestigious journals and conferences of the field, including PNAS, TBIOM, CVPR, ICCV, ECCV, FG, WACV, etc. He was a recipient of the ACM Multimedia Best Technical Full Paper Award in 2006.



I-Hong Jhuo is a senior applied scientist at Microsoft. He is an active participant in the development of innovative technologies for information retrieval and recommendation system while contributing to advances in the fields of computer vision, structured data and machine learning. Recognized by many awards, conducting the design of a top-performing video analytic system in NIST TRECVIDMED at Columbia University and ACM ACM Multimedia Grand Challenge 2012. His research interests include computer vision, information retrieval, and artificial intelligence.



Chia-Wen Lin (Fellow, IEEE) received his Ph.D. degree in Electrical Engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. Dr. Lin is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also a visiting professor with the Graduate School of Informatics, Kyoto University, Kyoto, Japan. His research interests include image/video processing and computer vision. He has served as a Fellow Evaluating Committee member (2021–2022), BoG Member-at-Large (2022–2024), and Distinguished Lecturer (2018–2019) of IEEE Circuits and Systems Society. He was Chair of IEEE ICME Steering Committee (2020–2021). He served as TPC Co-Chair of IEEE ICIP 2019 and IEEE ICME 2010, and General Co-Chair of IEEE VCIP 2018. He received two best paper awards from VCIP 2010 and 2015. He has served as an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and Neural Networks.

Yen-Yu Lin (Senior Member, IEEE) received the B.B.A. degree in Information Management, and the M.S. and Ph.D. degrees in Computer Science and Information Engineering from National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently a Distinguished Professor with the Department of Computer Science and Director of the Institute of Multimedia Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. His research interests include computer vision, machine learning, and artificial intelligence.