Enhancing Anchor-based Weakly Supervised Referring Expression Comprehension with Cross-Modality Attention

Ting-Yu Chu[1], Yong-Xiang Lin[1], Ching-Chun Huang[2], and Kai-Lung Hua[1,3]

[1] Dept. of CSIE, National Taiwan University of Science and Technology, Taiwan.
{m11115032,d10915006,hua}@mail.ntust.edu.tw
[2] Dept. of Computer Science, National Yang-Ming Chiao Tung University, Taiwan.
chingchun@cs.nycu.edu.tw
[3] Microsoft Taiwan.
kai.hua@microsoft.com

Abstract. Weakly supervised Referring Expression Comprehension (REC) tackles the challenge of identifying specific regions in an image based on textual descriptions without predefined mappings between the text and target objects during training. The primary obstacle lies in the misalignment between visual and textual features, often resulting in inaccurate bounding box predictions. To address this, we propose a novel cross-modality attention module (CMA) module that enhances the discriminative power of grid features and improves localization accuracy by harmonizing textual and visual features. To handle the noise from incorrect labels common in weak supervision, we also introduce a false negative suppression mechanism that uses intra-modal similarities as soft supervision signals. Extensive experiments conducted on four REC benchmark datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame. Our results show that our model consistently outperforms state-of-the-art methods in accuracy and generalizability. Our source code is publicly available at https://github.com/t22786959/Cross-Modality-Attention-in-weakly-supervised-REC.git

Keywords: Referring expression comprehension · cross-modality · similarity regularization · false negative suppression

1 Introduction

Referring Expression Comprehension (REC), often referred to as visual grounding, focuses on identifying specific objects in images based on descriptive expressions. Unlike traditional object detection, which works with predefined categories and relies solely on visual information, REC operates across both visual and textual modalities, allowing it to handle a broader range of applications in fields, such as robotics [12] and human-computer interaction [14, 26]. For instance, in robotics, REC can enhance object manipulation and navigation based on verbal

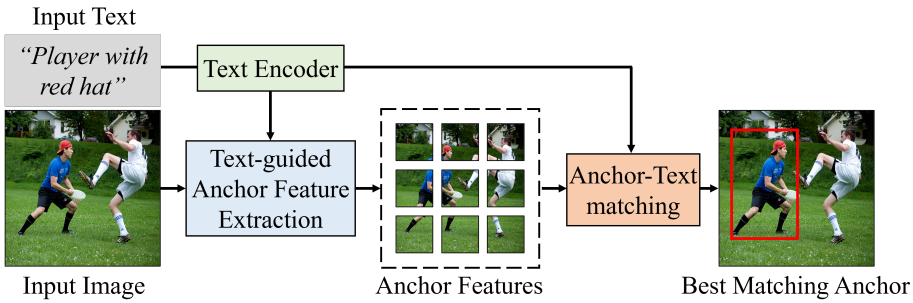

Fig. 1: Our anchor-based framework for weakly supervised referring expression comprehension. With the anchor features extracted from the text-guided anchor feature extration module, our model can obtain more discriminative features for the referred object.

instructions, while in human-computer interaction, it can improve user interfaces by enabling more natural and intuitive visual searches. Despite the progress made in fully supervised REC, the costly annotation process required at the instance level remains a significant obstacle to its advancement. This has led to increased interest in weakly supervised REC [10,22,24,32,36,39], which aims to reduce the reliance on detailed annotations. However, weakly supervised REC faces several key challenges. The primary obstacle lies in the misalignment between visual and textual features, often resulting in inaccurate bounding box predictions. Additionally, the absence of precise annotations makes it difficult to train models effectively, as they cannot rely on accurate instance-level labels for supervision.

To address the challenges of weakly supervised Referring Expression Comprehension, we propose a novel Cross-Modality Attention (CMA) module and the Feature-wise Linear Modulation (FiLM) module. The CMA module synchronizes text and visual features across different layers, while FiLM ensures precise alignment between specific text tokens and corresponding visual features at varying levels of abstraction. This layer-specific differentiation is essential for accurately capturing both low-level and high-level details. This approach significantly improves multi-layer localization, resulting in more accurate and robust bounding box predictions. In addition, we introduce an equivariant regularization method to tackle the lack of ground truth in weakly supervised contrastive learning. This method ensures the changes in semantics corresponds to consistent changes in similarity scores between anchor-text pairs. By aligning similarity scores between different anchor-text pairs, this approach compensates for the lack of precise annotations, providing more reliable supervision signals. This enhances the robustness and accuracy of feature representations, enabling effective model training even in the absence of accurate instance-level labels. Furthermore, to mitigate the noise introduced by false negative samples, we develop a mechanism that uses similarity scores between intra-modal pairs as soft supervision signals for inter-modal pairs. This ensures that inter-modal similarity

is consistent with intra-modal similarity, reducing the impact of false negatives and providing cleaner, more reliable training data.

Our initial framework, depicted in Fig 1, provides an overview of the overall architecture. However, these anchor-based approaches might not fully utilize the contextual information present in the entire scene, potentially limiting their effectiveness in more complex scenarios. Our approach builds upon recent advancements in one-stage detectors. Following previous work, we select YOLOv3 [30], which we adapt for weakly supervised REC. By leveraging the efficiency of one-stage detectors and addressing the challenges specific to weakly supervised learning, our method achieves state-of-the-art performance on four benchmark datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame. The main contributions of our work can be summarized as follows:

- **Enhance feature relevance and localization:** The Cross-Modality Attention (CMA) module improves the discriminative power of grid features and enhances localization accuracy by optimizing their relevance within the textual context.
- **Strengthen multi-modal similarity alignment:** Our equivariant regularization technique enhances the alignment between multi-modal(visual and textual) similarities, ensuring they accurately reflect semantic changes.
- **Suppress noise from false negative samples:** We develop a mechanism that uses intra-modal similarities as soft supervision signals, reducing noise and improving training reliability.
- **State-of-the-art Performance**: Extensive evaluation on four REC benchmark datasets (RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame) demonstrates that our approach consistently surpasses existing state-of-the-art methods in both accuracy and generalizability, highlighting its robust performance across diverse scenarios.

2 Related Work

2.1 Weakly Supervised Referring Expression Comprehension

In weakly supervised Referring Expression Comprehension (REC), the annotation process is both time-consuming and labor-intensive. In recent years, the weakly supervised training scheme has increasingly gained attention. In contrast to fully supervised REC, weakly supervised REC presents greater challenges due to the absence of bounding box annotations. Many existing techniques are developed using two-stage supervised REC models [24,36] and frame weakly supervised REC as a region-text ranking problem. The primary challenge lies in providing effective supervision signals from image-text pairs. To tackle this challenge, current approaches utilize methodologies such as sentence reconstruction [22,32] and contrastive learning [10,39]. These approaches work by constructing both positive and negative sample pairs from carefully selected regions and expressions, and then calculating the contrastive loss. However, all

paradigms ignore the heterogeneous gap between textual descriptions and visual images. In this paper, we utilize a contrastive learning-based approach and introduce the CMA module to obtain higher-quality sample pairs.

2.2 Equivariant Similarity

Equivariant Similarity regularization, aimed at faithfully reflecting the semantic change between different image-text pairs. In other words, the same semantic changes should lead to a similar amount of similarity changes. The equivariance property plays an important role in various fields, including self-supervised learning [6,38], representation learning [29], and language understanding [9].

In order to focus on this issue, [37] points out the significance of the equivariant similarity measure in vision-language models, imposing additional equivariance regularization on image-text pairs for vision-language foundation models learning without additional supervision. Based on its exploration on equivariant similarity, we propose a novel loss for the regularization of equivariance.

2.3 False Negatives in Contrastive Learning

In the conventional methodology of contrastive learning, the generation of positive sample pairs is typically achieved through the application of data augmentation techniques. These techniques involve the manipulation of original data samples in order to create new, but similar samples. Meanwhile, the rest of the samples within the same batch are defined and treated as negative sample pairs, regardless of semantic similarities. This scheme inevitably encounters the false negatives issue, where instances sharing identical semantic concepts are incorrectly labeled as negatives, leading to misguidance in model learning.

Based on this, some works use clustering-based methods [18] to encode semantic structures and then perform contrastive learning on these semantically similar cluster centers. Some research focuses on improving architectures [5] or data augmentation [34] strategies. Despite these advances, there is little work addressing the false negative problem directly. This problem can significantly impact the performance and reliability of contrastive learning models.

3 Methodology

3.1 Overview

Referring Expression Comprehension (REC) aims to locate the target instance within an image by generating a bounding box, guided by a provided text expression. In the current weakly supervised setting, it is challenging for the model to acquire detection bounding box merely on text expressions and images. In this case, existing weakly supervised solutions usually adopt a pre-trained object detection network, to provide a set of candidate bounding boxes. Afterwards, the model conducts weakly supervised training based on semantic reconstruction [23,32] or cross-modal contrastive learning [11,40]. As shown in Fig 2, our

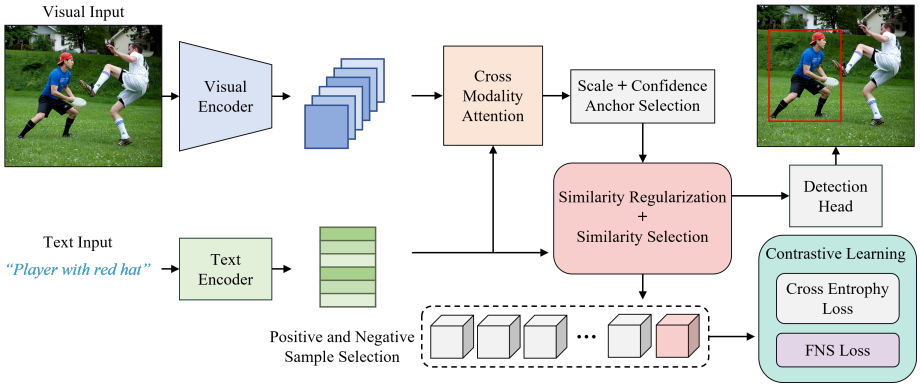

Fig. 2: Illustrates the overall architecture of our network, adapted from RefCLIP [15]. It includes a novel cross-modality attention module and a similarity regularization method. Additionally, false negative suppression is implemented to enhance sample quality.

architecture is adapted from the model RefCLIP [15]. The original RefCLIP framework employs the efficient one-stage detector YOLOv3 [30] to build the visual encoder. The language encoder is a bidirectional GRU [3] followed by a self-attention layer [35]. Given an image I and a text expression T, along with a visual encoder E_v and a text encoder E_t, RefCLIP also simplifies the REC task to an anchor-text matching problem, i.e., which anchor is most likely to have the target box:

$$a^* = \arg\max_{a \in A} \phi(E_v(I), E_t(T), a), \tag{1}$$

where a^* represents the optimal anchor, with A representing the set of anchor points in image I provided by the pretrained YOLOv3, and $\phi(\cdot)$ denoting a simple linear ranking module like cosine similarity.

Our overall model architecture is as shown in Fig 2. Our proposed architecture introduces an additional cross-modal attention module (CMA). The feature maps extracted by the backbone of pretrained YOLOv3 and the language features extracted by the bidirectional GRU, denoted as Y, will then be fed into the cross-modal attention module. By integrating textual features into image features, the proposed approach emphasizes the areas expressed in the text within the image features. This enhances the accuracy of weakly-supervised contrastive learning by fusing the text feature information with the image features.

3.2 Cross-Modality Attention Module

To enhance the relevance of extracted image features to the described objects in text, our objective is to optimize the discriminative of these features within the context of textual input. Our approach involves a detailed sequence of operations at each layer of the CMA module, as shown in Fig 3, aimed at achieving this goal.

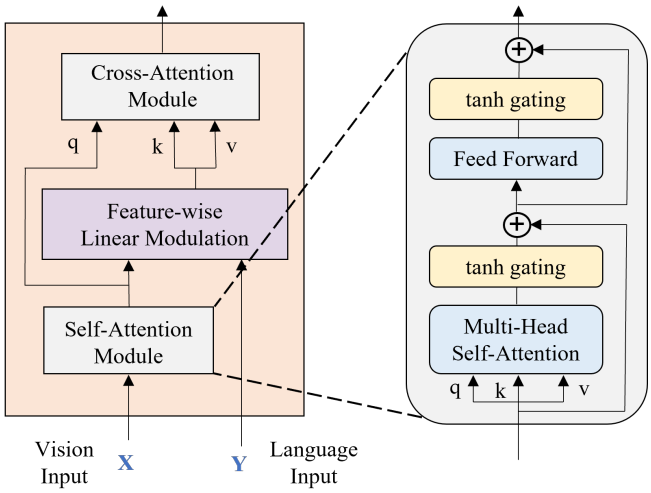

Fig. 3: A illustration of the architecture of the Cross-Modality Attention Module. It consists of three key components: a self-attention layer, which preserves the original visual features; Feature-wise Linear Modulation, responsible for modulating the features in a feature-wise manner; and a cross-attention layer, which facilitates effective information exchange between image and text modalities by fusing the cross-modality features.

Firstly, we start by applying a layer of multi-head self-attention [19,31] to the raw image features, denoted as $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels respectively. The self-attention mechanism computes attention weights for each pair of feature vectors, capturing the importance of spatial relationships within the image. The input data X is transformed into three distinct representations: $Query, Key$, and $Value$. These transformations are achieved through the multiplication of X with corresponding weight matrices W_q, W_k, and W_v, where W_q, W_k, and W_v are learnable weight matrices. The attention weights are computed using the scaled dot-product attention mechanism:

$$F = \text{softmax}(\frac{Query \cdot Key^T}{\sqrt{C}}) \cdot Value \tag{2}$$

Then, the computed attention feature map F is fed into a dense feed-forward layers, which is composed of 2 linear layer and 1 gelu layer. Additionally, inspired by [1], we insert tanh-gating layer after self-attention layer and feed-forward layer respectively. The tanh-gating mechanism involves multiplying the output of a newly introduced layer by $\tanh(\alpha)$ before combining it with the input representation from the residual connection. Here, α represents a layer-specific learnable scalar initialized to 0 [2]. This approach allows our model output to match that of the pretrained YOLOv3 at initialization, improving training stability and final

performance. This $\tanh(\alpha)$ layer ensures a stable starting point, mitigates early training fluctuations, and provides flexibility for gradual optimization.

Continuing with our methodology, we introduce a Feature-wise Linear Modulation (FiLM) module modified from [13] into our framework. This module dynamically amplifies fine-grained features at different layers. This differentiation across layers is essential as it emphasizes the alignment between specific text tokens and corresponding visual features at varying levels of abstraction, ensuring that both low-level and high-level details are accurately captured and matched. Given the generated language feature Y and the attention map F obtained in the previous stage. The specific operations are as follows:

$$\gamma = \mathrm{MLP}_\gamma(Y), \quad \beta = \mathrm{MLP}_\beta(Y), \tag{3}$$

$$F' = \mathrm{ReLU}\left(\gamma \odot F \oplus \beta\right), \tag{4}$$

where MLP_γ and MLP_β are two one-layer MLPs that map language vector Y to coefficients γ and β. Then we apply these coefficients to visual feature F followed by a ReLU operations, yielding the output $F' \in \mathbb{R}^{H\times W\times C}$, where $\odot$ and $\oplus$ represent the broadcast element-wise multiplication and addition, respectively. This step enhances the discriminative of the extracted features by incorporating semantic information from the textual input.

Furthermore, our model architecture incorporates a cross-attention mechanism, which facilitates a more cohesive integration of textual and visual information. This mechanism operates by attending to both modalities simultaneously, enabling the model to construct a collaborative relationship between text and image features. The definitions of $Query', Key'$, and $Value'$ among different modalities are as follows:

$$\begin{aligned} Query' &= F \cdot W_q \\ Key' &= F' \cdot W_k \\ Value' &= F' \cdot W_v \end{aligned} \tag{5}$$

Then the cross attention weights are computed using the scaled dot-product attention mechanism. Our cross attention mechanism design involves using the attention map F, produced by the self-attention layer, as the $Query'$. This strategic choice enables the model to focus on the most relevant regions within the image. Additionally, we utilize the image feature F' generated by the Feature-wise Linear Modulation, serving it as both Key' and $Value'$. This integration facilitates a thorough understanding of the image content by establishing associations between textual descriptions and visual features. By leveraging the complementary nature of text and images, this layer promotes a more stable and nuanced fusion of the two modalities.

By integrating the CMA module and the FiLM module. Our model ensures precise alignment between specific text tokens and corresponding visual features at varying levels of abstraction. This layer-specific differentiation is essential for accurately capturing both low-level and high-level details, which are often overlooked in existing methods like RefCLIP [15].

3.3 Similarity Regularization

Equivariant similarity refers to the conception that changes in semantics should correspond to consistent changes in similarity scores between anchor-text pairs. In other words, when the meaning of an anchor and its text changes, the similarity between them should change consistently with the meaning changes. To address this similarity regularization issue, some researchers, like Cyclip [8], focus on regularizing the CLIP cosine similarity score, S, with intra-modal consistency, i.e., forcing $S(A_1, A_2)$ is close to $S(T_1, T_2)$ and cross-modal consistency, i.e., forcing $S(A_1, T_2)$ is close to $S(A_2, T_1)$. Unlike these previous studies, we focus on ensuring that changes in similarity scores should faithfully respect to the semantic changes. We propose a similarity regularization loss. We define A_1 as the anchor feature that best matches the text feature T_1 in cosine similarity, and A_2 as the anchor feature that best matches the text feature T_2 in cosine similarity. Using the two matched anchor-text pairs $\{A_i, T_i\}$ and $\{A_j, T_j\}$, we can calculate four similarity scores: S_{ii}, S_{ij}, S_{jj}, and S_{ji}. Here, S_{ii} represents the similarity score between anchor feature A_i and text feature T_i. Similarly S_{ij} denotes the similarity score between anchor feature A_i and text feature T_j. In our implementation, we adopt mean average error to regularize the equation of similarities and utilize a margin parameter ω to control the strength of regularization. For any two matched anchor-text pairs $\{S_{ii}$,$S_{jj}\}$ in a batch, our

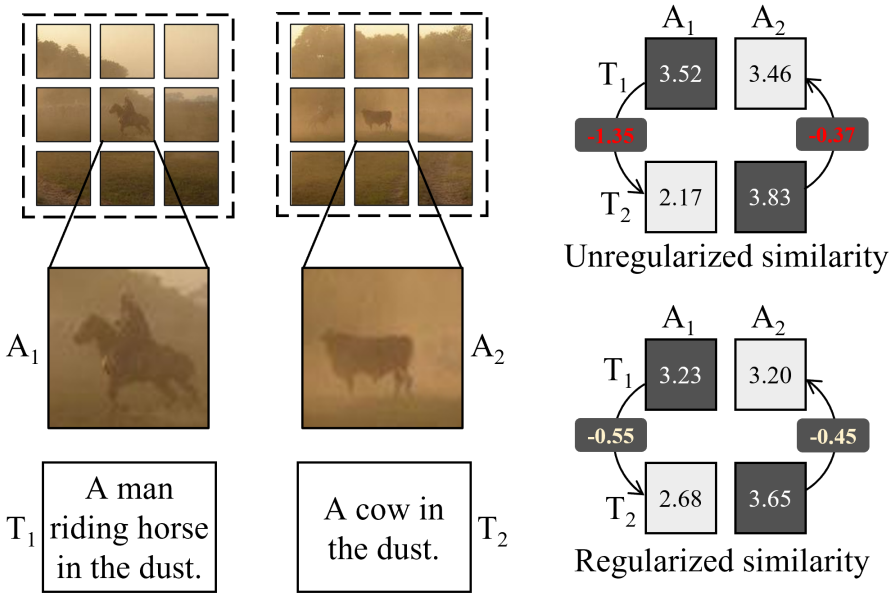

Fig. 4: Measuring the similarity score change of unregularized similarity and regularized similarity.

optimization objective is formulated as:

$$\begin{aligned} L_{regular}(i,j) = & L_1\left(S_{ii} - S_{ij}, S_{jj} - S_{ji}\right) \\ & + L_1\left(S_{ii} - S_{ji}, S_{jj} - S_{ij}\right) - \omega \end{aligned} \tag{6}$$

where L_1 represents the L_1 distance. Then we adopt mean average error to calulate the loss of similarity regularization. By the design of our similarity regularization loss, our model is capable of more accurately reflecting semantic changes in similarity scores. This enhancement allows for a deeper understanding of the evolving semantic context within the model's predictions.

Fig 4 illustrates the effectiveness of equivariant similarity. We consider two matched anchor-text pairs denoted as $\{A_1, T_1\}$ and $\{A_2, T_2\}$ that are semantically similar but only different in the object in the dust. One of the objects is a cow, and the other is a man riding a horse. The changes in similarity scores guided by the unregularized similarity are highly inconsistent (-1.35 v.s. -0.37). However, our regularized similarity between anchor and text accurately captures semantic changes, ensuring that equivalent semantic alterations correctly result in comparable changes in similarity scores. For example, as illustrated in Fig 4, the same semantic changes lead to a similar amount of change (-0.55 v.s. -0.45) in similarity scores in our regularized approach.

3.4 False Negative Suppression

In weakly supervised contrastive learning, negative anchor-text pairs are created using other anchors within the same batch. Without the supervision of labels, the selected negative pairs could actually belong to the same semantic category. We define these undesirable negatives as false negatives, i.e., negative pairs from the same semantic category. According to a study from [33], the false negative pairs existed in every batch that may harm the learning process. To suppress the false negative effects, we introduce a false negative suppresssion loss. An overall architecture is shown as Fig 5. We use the similarity scores between intra-modal pairs as a soft supervision signal for inter-modal pairs to mitigate the impact of false negative samples. Put simply, when two anchor scenes in the same batch are semantically similar, their corresponding text expressions should also be similar. Based on this theory, we hope that the inter-modal similarity should be consistent with the intra-modal similarity.

In our implementation, we employ the anchor features $Z^a \in \mathbb{R}^{B\times H'\times W'\times C'}$ and text features $Z^t \in \mathbb{R}^{B\times C'}$ extracted through the cross-modality attention module, where B, H', W', and C' represent the batch size, height, width, and number of channels of the anchor point, respectively. Then we use Z^t to calculate a dot product with $(Z^t)^T$ to obtain the pairwise self-similarity matrix $W^t \in \mathbb{R}^{B\times B}$ for text-modal. We denote W^t_{ij} as the element in the i-th row and j-th column of W^t. Likewise, we process the same operations to the anchor features and obtain a visual adjacency matrix $W^a \in \mathbb{R}^{B\times B}$. S_{ij} represents the inter-modal similarity between A_i and T_j. For example, when the negative pairs S_{ij} are similar, the similarity value of their text features, W^t_{ij}, should be high, as well as

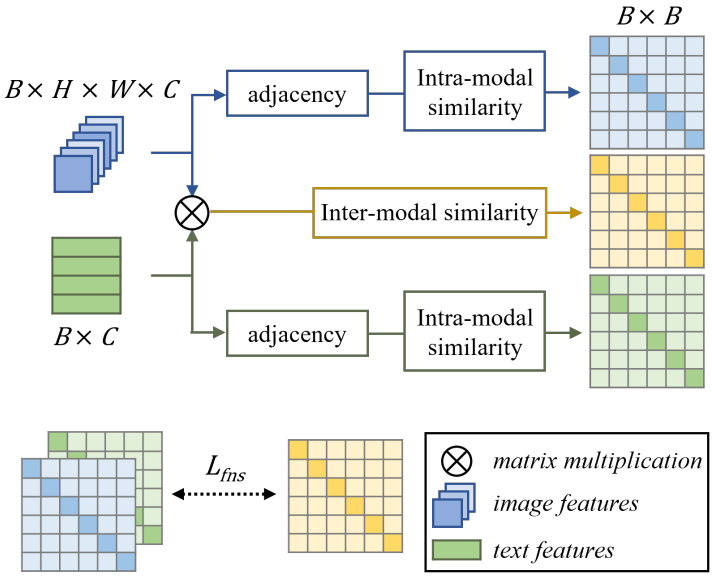

Fig. 5: An overall architecture of False Negative Suppression. The image adjacency matrix and text adjacency matrix are respectively constructed to calculate the intra-modal similarity.

the visual similarity, W_{ij}^{a}. And we observed that the similarity scores between different modalities exhibit a heterogeneous gap. Therefore, we use the average similarity score between the textual and visual modalities as the soft supervision signal. Then we implement a loss function to mitigate the impact of false negative samples on misleading the model's training process. For a sample S_{ij} in the inter-modal similarity matrix, its optimization objective is formulated as:

$$L_{fns}(i,j) = L_1\left(S_{ij}, \frac{1}{2}\left(W_{ij}^{a} + W_{ij}^{t}\right)\right) \tag{7}$$

Specifically, we adopt mean average error to calulate the loss of false negative suppression. We compute the L_1 distances between the inter-modal contrastive matrix and intra-modal adjacency matrices, thereby accounting for all pairwise similarities across modalities.

In the overall formulation of the loss function, we incorporate the similarity regularization method and the false negative suppression method into our model. Both methods can be seamlessly integrated with our contrastive learning framework as additional optimization terms. In practice, we introduce two learnable balancing factors, denoted as λ and λ'. These balancing factors play a crucial role in ensuring that our model achieves the desired balance between different objectives or constraints. By allowing the model to automatically adjust these factors, it can acquire more valuable learning directions during the

training process. As a result, the final loss can be as:

$$Loss_{total} = L_{contrastive} + \lambda L_{regular} + \lambda' L_{fns} \tag{8}$$

where $L_{Contrastive}$ represents the basic contrastive loss, and λ and λ' are learnable parameters. And the contrastive loss $L_{Contrastive}$ used in our work is defined as follows:

$$L_{contrastive} = -\log \frac{\exp\left(\text{sim}(f_{a_0}^i, f_t^i)/\tau\right)}{\sum_{n=0}^{N}\sum_{j=0}^{M} \mathbb{I}_{\neg(i=j \wedge n \neq 0)} \exp\left(\text{sim}(f_{a_n}^j, f_t^i)/\tau\right)} \tag{9}$$

where $f_{a_0}^i$ is the feature of the correct anchor in the i-th image, $f_{a_n}^j$ represents the features of other negative anchors, $\text{sim}(f_a, f_t)$ denotes the similarity between anchor f_a and text feature f_t, τ is the temperature scaling parameter, $\mathbb{I}$ is an indicator function to filter out the positive pair, and the sums over N and M account for negative anchors across images. This loss function effectively maximizes the alignment of correct anchor-text pairs while minimizing the similarity of incorrect pairs, ensuring robust cross-modal matching.

4 Experiments

4.1 Datasets and Metric

RefCOCO [27] comprises 142,210 referring expressions and 50,000 objects extracted from 19,994 images sourced from the MSCOCO dataset [20], which is divided into four sets: train with 120,624 expressions, validation with 10,834 expressions, testA with 5,657 expressions, and testB with 5,095 expressions. The referring expressions in RefCOCO are primarily about absolute spatial information. **RefCOCO+** [27] contains 141,564 referring expressions corresponding to 49,856 bounding boxes extracted from 19,992 MSCOCO images. The data splits are divided into train with 120,191 referring expressions, validation with 10,758 referring expressions, testA with 5,726 referring expressions, and testB with 4,889 referring expressions. However, the descriptions within RefCOCO+ primarily focus on relative spatial information and visual attributes such as color and texture. **RefCOCOg** [25, 27] consists of 104,560 referring expressions associated with 54,822 bounding boxes found in 26,711 images. In comparison to RefCOCO and RefCOCO+, the referring expressions found in RefCOCOg typically exhibit greater length and complexity. For our experiments, we utilize the Google split [25] of RefCOCOg. **ReferItGame** [16] comprises 19,997 images sourced from the SAIAPR-12 dataset, featuring 99,220 bounding boxes and 120,072 referring expressions. We partition the dataset into training, validation, and test sets according to the Berkeley split. Our evaluation metric is **IoU@0.5**, where a prediction is considered correct if the Intersection over Union (IoU) between the predicted bounding box and the ground-truth box exceeds 0.5.

Table 1: Comparisons with state-of-the-art methods on RefCOCO [27], RefCOCO+ [27], RefCOCOg [25,27], ReferItGame [16]. We conduct comprehensive comparisons across four benchmark datasets specifically designed for REC. These experiments enable us to evaluate and measure the effectiveness and accuracy of our approach in a systematic and comprehensive manner.

Method	RefCOCO			RefCOCO+			RefCOCOg	ReferItGame
	val	testA	testB	val	testA	testB	val-g	test
VC [28]	-	32.68	27.22	-	34.68	28.10	29.65	14.50
KAC Net [4]	-	-	-	-	-	-	-	15.83
MATN [41]	-	-	-	-	-	-	-	13.61
ARN [22]	32.17	35.25	30.28	32.78	34.35	32.13	33.09	26.19
IGN [40]	34.78	37.64	32.59	34.29	36.91	33.56	34.92	-
DTWREG [32]	38.35	39.51	37.01	38.91	39.91	37.09	42.54	-
RelR [24]	-	-	-	-	-	-	-	37.68
NCE+Distillation [36]	-	-	-	-	-	-	-	38.39
RefCLIP [15]	60.36	58.58	57.13	40.39	40.45	38.86	47.87	**39.58**
Ours	**63.91**	**61.44**	**63.76**	**41.82**	**42.05**	**40.37**	**48.38**	38.83

4.2 Implementation Details

Our model was implemented using PyTorch and trained on a single RTX 3090 GPU. We conducted the experiment following all the training and evaluation settings outlined in RefCLIP [15]. During training, our proposed model is trained over 25 epochs using the Adam optimizer [7], with a batch size of 64. The input image is resized to 416 ÃŮ 416, and the anchor features are projected into 512 dimensions through multi-scale fusion. The language encoder's dimension is set to 512, and the maximum length of the input text is restricted to 15 for RefCOCO, RefCOCO+, and RefCOCOg, and to 20 for ReferItGame. We then use a pretrained YOLOv3 [30], with the DarkNet-53 backbone, as the detector to extract anchor features. This YOLOv3 model, pretrained on MS-COCO [20], excludes images from the validation and test sets across the three datasets mentioned earlier in section 4.1. For a fair comparison with [21,36] in ReferItGame, we use the YOLOv3 pretrained on Visual Genome [17] as the detector of our model. The pretrained YOLOv3 is frozen throughout the entire training process.

4.3 Results

In Table 1, we compare our proposed model with common weakly supervised REC models, including both one-stage and two-stage REC models, across four different datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame. Our research shows that our model outperforms these methods significantly. Notably, we achieve accuracies of 63.91%, 61.44% and 63.76% on the val, testA and testB splits of RefCOCO, which bring absolute improvements of 3.55, 2.86 and 6.63 percentage points respectively over the previous best performance of the anchor-based one-stage framework RefCLIP [15]. For the RefCOCO+ dataset, our model

achieve accuracies of 41.82%, 42.05% and 40.37% on the val, testA and testB splits, corresponding to absolute improvements of 1.43, 1.60 and 1.51 percentage points. Our model's superior performance extends beyond RefCOCO and RefCOCO+ to include RefCOCOg and ReferItGame datasets, establishing it as the current top-performing method. The visualizations in Fig 6 demonstrate the effectivity of our approaches. Specifically, the visualizations highlight the improvements in bounding boxes accuracy by comparing our model's predictions with the ground truth and with predictions from baseline model.

Table 2: Evaluation of each proposed module on RefCOCO and RefCOCOg.

Method	RefCOCO			RefCOCOg
	val	testA	testB	val-g
Baseline	60.36	58.58	57.13	47.87
C	62.27	59.95	60.07	<u>48.12</u>
C + E	<u>63.16</u>	<u>60.80</u>	<u>62.66</u>	48.06
C + E + F	**63.91**	**61.44**	**63.76**	**48.38**

Table 3: Evaluation of each proposed module on RefCOCO+ and ReferItGame.

Method	RefCOCO+			ReferItGame
	val	testA	testB	test
Baseline	40.39	40.45	38.86	**39.58**
C	41.02	41.58	39.65	38.70
C + E	<u>41.35</u>	<u>41.82</u>	<u>40.02</u>	38.74
C + E + F	**41.82**	**42.05**	**40.37**	<u>38.83</u>

4.4 Ablation Study

In Table 2 and Table 3, specific symbols represent the modules: C denotes the use of the cross-modality attention module, E signifies similarity regularization, and F indicates false negative suppression. These table allow us to analyze and understand the impact of each module on the performance of accuracy IoU@0.5 on our proposed model. In Table 2, the addition of each module to the model results in an average 0.9% improvement in accuracy scores across all splits of the RefCOCO dataset, highlighting the effectiveness of each component. In Table 3, our proposed modules show consistent improvements across the RefCOCO+ dataset, with noticeable gains in each splits. However, these enhancements are not observed in the ReferItGame dataset, this can be attributed to its unique

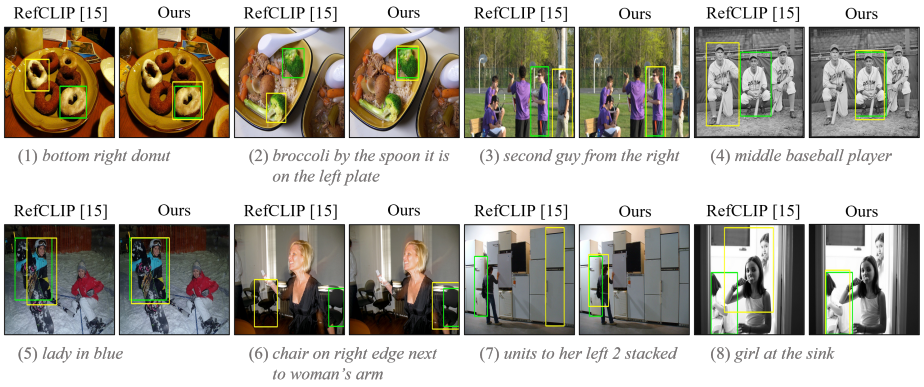

Fig. 6: The visualization results of our model compared to the baseline model RefCLIP [15]. The yellow bounding boxes represent the predicted results, while the green bounding boxes denote the ground truth. Sub-figure 1 to 4 demonstrate that our proposed cross-modality attention module help the model get more accurate bounding boxes. Sub-figure 5 to 8 illustrate that our proposed similarity regularization and false negative suppression approaches enhance the model's ability to generate more precise bounding boxes.

characteristics. Unlike other datsets, ReferItGame's images primarily consist of natural landscapes, which lack easily recognizable features, making object detection more difficult. By individually incorporating different modules into our model, we can observe that each module we proposed contributes to an overall improvement in the model's performance.

5 Conclusion

In conclusion, our study addresses the critical challenge of aligning image and text features in weakly supervised REC. By introducing the cross-modality attention module, we effectively bridge the gap between visual and textual modalities, mitigating the discrepancies that often lead to incorrect bounding box predictions. Our incorporation of equivariant regularization based on similarity further enhances cross-modal alignment by accurately reflecting semantic changes. Additionally, we introduce the mechanism for false negative suppression tackles the noise in weakly supervised contrastive learning, improving the robustness of the model without requiring additional ground truth supervision. The extensive experiments conducted on the RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame datasets validate the effectiveness of our approaches. Our results consistently outperform state-of-the-art methods, demonstrating superior accuracy and generalizability. This work not only advances the REC field but also provides a strong foundation for future research in cross-modal understanding and weakly supervised learning.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bachlechner, T., Majumder, B.P., Mao, H., Cottrell, G., McAuley, J.: Rezero is all you need: Fast convergence at large depth. In: Uncertainty in Artificial Intelligence. pp. 1352–1361. PMLR (2021)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Chen, K., Gao, J., Nevatia, R.: Knowledge aided consistency for weakly supervised phrase grounding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4042–4050 (2018)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., SoljaÄĐiÄĞ, M.: Equivariant contrastive learning. arXiv preprint arXiv:2111.00899 (2021)
7. Diederik, P.K.: Adam: A method for stochastic optimization. (No Title) (2014)
8. Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., Grover, A.: Cyclip: Cyclic contrastive language-image pretraining. Advances in Neural Information Processing Systems **35**, 6704–6719 (2022)
9. Gordon, J., Lopez-Paz, D., Baroni, M., Bouchacourt, D.: Permutation equivariant models for compositional generalization in language. In: International Conference on Learning Representations (2019)
10. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: European Conference on Computer Vision. pp. 752–768. Springer (2020)
11. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: European Conference on Computer Vision. pp. 752–768. Springer (2020)
12. Hsia, H.A., Lin, C.H., Kung, B.H., Chen, J.T., Tan, D.S., Chen, J.C., Hua, K.L.: Clipcam: A simple baseline for zero-shot text-guided object and action localization. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4453–4457. IEEE (2022)
13. Huang, B., Lian, D., Luo, W., Gao, S.: Look before you leap: Learning landmark features for one-stage visual grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16888–16897 (2021)
14. Huang, W.S., Hong, B.K., Cheng, W.H., Sun, S.W., Hua, K.L.: A cloud-based intelligent skin and scalp analysis system. In: 2018 IEEE Visual Communications and Image Processing (VCIP). pp. 1–5. IEEE (2018)
15. Jin, L., Luo, G., Zhou, Y., Sun, X., Jiang, G., Shu, A., Ji, R.: Refclip: A universal teacher for weakly supervised referring expression comprehension. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2681–2690 (2023)
16. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)

17. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
18. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966 (2020)
19. Lin, H.H., Lin, J.D., Ople, J.J.M., Chen, J.C., Hua, K.L.: Social media popularity prediction based on multi-modal self-attention mechanisms. IEEE Access **10**, 4448–4455 (2021)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
21. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1950–1959 (2019)
22. Liu, X., Li, L., Wang, S., Zha, Z.J., Meng, D., Huang, Q.: Adaptive reconstruction network for weakly supervised referring expression grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2611–2620 (2019)
23. Liu, X., Li, L., Wang, S., Zha, Z.J., Su, L., Huang, Q.: Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 539–547 (2019)
24. Liu, Y., Wan, B., Ma, L., He, X.: Relation-aware instance refinement for weakly supervised visual grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5612–5621 (2021)
25. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
26. Miao, L., Chen, S.F., Hsu, Y.L., Hua, K.L.: How does c-v2x help autonomous driving to avoid accidents? Sensors **22**(2), 686 (2022)
27. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 792–807. Springer (2016)
28. Niu, Y., Zhang, H., Lu, Z., Chang, S.F.: Variational context: Exploiting visual and textual context for grounding referring expressions. IEEE transactions on pattern analysis and machine intelligence **43**(1), 347–359 (2019)
29. Qi, G.J., Zhang, L., Lin, F., Wang, X.: Learning generalized transformation equivariant representations via autoencoding transformations. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 2045–2057 (2020)
30. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
31. Shahid, M., Virtusio, J.J., Wu, Y.H., Chen, Y.Y., Tanveer, M., Muhammad, K., Hua, K.L.: Spatio-temporal self-attention network for fire detection and segmentation in video surveillance. IEEE Access **10**, 1259–1275 (2021)
32. Sun, M., Xiao, J., Lim, E.G., Liu, S., Goulermas, J.Y.: Discriminative triad matching and reconstruction for weakly referring expression grounding. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4189–4195 (2021)

33. Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., Barnes, N.: Learning audio-visual source localization via false negative aware contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6420–6429 (2023)
34. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in neural information processing systems **33**, 6827–6839 (2020)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Å., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
36. Wang, L., Huang, J., Li, Y., Xu, K., Yang, Z., Yu, D.: Improving weakly supervised visual grounding by contrastive knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14090–14100 (2021)
37. Wang, T., Lin, K., Li, L., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Equivariant similarity for vision-language foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11998–12008 (2023)
38. Xie, Y., Wen, J., Lau, K.W., Rehman, Y.A.U., Shen, J.: What should be equivariant in self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4111–4120 (2022)
39. Zhang, Z., Zhao, Z., Lin, Z., He, X.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems **33**, 18123–18134 (2020)
40. Zhang, Z., Zhao, Z., Lin, Z., He, X., et al.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems **33**, 18123–18134 (2020)
41. Zhao, F., Li, J., Zhao, J., Feng, J.: Weakly supervised phrase localization with multi-scale anchored transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5696–5705 (2018)