DiffusionREC: Diffusion Model with Adaptive Condition for Referring Expression Comprehension

Jingcheng Ke[1], Waikeung Wong[2*], Jia Wang[3], Mu Li[4], Lunke Fei[1], Jie Wen[4]

[1]School of Computer Science and Technology, Guangdong University of Technology, GuangZhou, China
[2]School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong
[3]College of Medical Information Engineering, Guangdong Pharmaceutical University
[4]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
freedom6927@gmail.com, calvin.wong@polyu.edu.hk, {jiawangmq, limuhit}@gmail.com, {flksxm, jiewen_pr}@126.com

Abstract

The objective of referring expression comprehension (REC) is to accurately identify the object in an image described by a given expression. Existing REC methods, including transformer-based and graph-based approaches among others, have shown robust performance in REC tasks. In this study, we present a groundbreaking framework named DiffusionREC for REC task. This framework reimagines the REC task as a text-guided bounding box denoising diffusion process, through which noisy bounding boxes are refined and distilled to pinpoint the target box. Throughout the training process, the bounding box of the target object diffuses from its ground-truth position towards a random distribution. Simultaneously, a filtering-based object decoder is introduced to reverse this diffusion of noise, conditional on the provided expression, the result from previous denoised step and the interaction between the expression and the image. At the inference stage, we begin by randomly generating a collection of boxes. Subsequently, the filtering-based object decoder is iteratively employed to refine and prune these bounding boxes, taking into account aforementioned conditions. Extensive experiments on five datasets show that DiffusionREC not only effectively addresses the limitations of existing REC methods but also surpasses them, delivering superior performance.

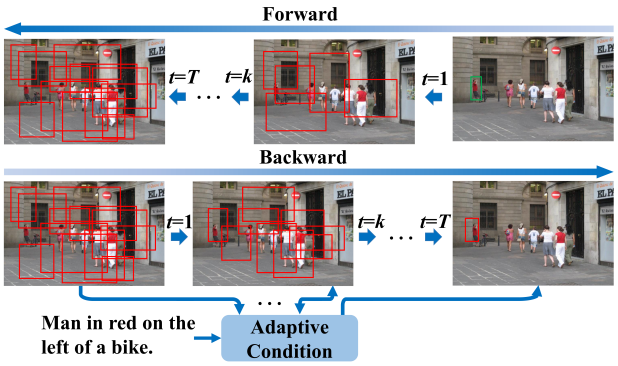

Figure 1: The operation of DiffusionREC involves a forward and backward process. In the forward phase, bounding boxes are progressively added to the image until they completely encompass it. During the backward phase, the expression, vision-language features from our visual-linguistic transformer encoder, and previous results are combined and fed into the adaptive condition module for fine-tuning. This refined condition then steers the decoder to effectively adjust the positions of bounding boxes and eliminate irrelevant bounding boxes.

Introduction

Artificial intelligence technologies are extensively utilized in real-world applications (Li et al. 2024a,b; Yang, Che, and Leung 2025; Jia-Mu Sun and Gao 2024). Among these applications, joint vision-language representation learning has emerged as a dynamic research area within computer vision, offering numerous applications such as referring expression comprehension (REC) (Hamilton et al. 2024; Zhang, Luo, and Lei 2024), referring expression generation (REG) (Sun et al. 2023a), referring expression segmentation (RES) (Liang et al. 2022), visual question answering (Dancette et al. 2023), image captioning (Luo et al. 2023), tracking (Tianyang Xu and Wu 2023) and scene understanding (Peng et al. 2023).

In these facets, REC plays a pivotal and integrative role in the joint visual-textual interpretation of a target object identified by an accompanying expression. REC aims to precisely locate the object described by the expression. For example, when provided with the expression "man in blue on the leftmost side," the goal of REC is to accurately localize the said "man in blue", ensuring that his position is consistent with the textual description in the associated image or video (Fang et al. 2024b,a).

Although recent state-of-the-art REC methods have shown improvements in performance, they still face two notable issues. Firstly, some of them (i.e. transformer-based REC methods) depend on a set of predetermined learnable queries. These approaches often face challenges in achieving convergence because they require attention on a specific region throughout the entire image. Furthermore, these methods are sub-optimal when dealing with small-sized target objects. Secondly, the remainder (i.e., two-stage REC methods) rely on the performance of the detector. If the detector fails to accurately identify the location of the object, these methods cannot correctly localize the intended object. Re-

*Corresponding Author: Waikeung Wong.

cently, some scholars (Chen et al. 2023) introduced a new framework called DiffusionDet for object detection, which formulates object detection as a denoising diffusion process from noisy boxes to object boxes. Considering the parallels between object detection and REC, it is logical to pose the following question: Is it possible to develop an effective diffusion-based framework for REC that would resolve the aforementioned challenges?

Inspired by the work presented in (Chen et al. 2023), as illustrated in Fig 1, we introduce an innovative framework named DiffusionREC. Our approach begins with the deployment of a unique visual-linguistic transformer encoder, utilizing prompt learning to integrate the given image and textual expression. Subsequently, We generate a random set of noisy bounding boxes and input them into a filtering-based object decoder for iterative decoding. This process is conditioned on the expression, the vision-language features from our visual-linguistic transformer encoder, and the output from the previous denoising step. In other words, with each step of decoding, the conditional input is dynamically adjusted based on the previous output from the decoder. The process culminates in identifying the decoded bounding box that corresponds to the ground-truth box, which is then designated as the definitive prediction.

Historically, some visual-linguistic transformer encoders (Ye et al. 2022) have focused on emphasizing objects in images during visual feature embedding while often overlooking the importance of specific words during text feature embedding. Recognizing that not all words in an expression are equally important—particularly nouns and relational terms—we propose a novel visual-linguistic transformer encoder to extract key information from both image and expression. This approach enhances vision-language features extraction, leading to improved the performances of REC methods.

Diffusion models have recorded remarkable achievements across a variety of generative tasks and are beginning to be utilized in perceptual tasks, such as image segmentation and object detection. Yet, as far as we are aware, there has been no successful application of this model in REC. Our DiffusionREC goes beyond merely eliminating the constraints imposed by detectors or a fixed set of learnable queries, it also demonstrates exceptional flexibility. Experimental results on RefCOCO, RefCOCO+, RefCOCOg, Flickr30K and RefClef datasets demonstrate that our DiffusionREC outperforms existing approaches, achieving state-of-the-art (SoTA) performances.

In summary, there are three contributions in this paper.

- To the best of our knowledge, we are the first to apply diffusion models to REC tasks. Using a filtering-based object decoder that processes the expression and previous decoding results, we efficiently refine and remove noisy bounding boxes during the reverse diffusion process. This greatly enhances the model's accuracy in identifying the target object described in the expression.
- We introduce a novel visual-linguistic transformer encoder that focuses on the critical informational components embedded within the visual tokens of the image and the text tokens of the expression.
- We validate our assertions through extensive evaluations across five datasets, illustrating that our DiffusionREC showcases exceptional performances.

Related Works

Referring Expression Comprehension (REC)

In this subsection, we briefly review the recent developments in various REC methods, categorizing them into one-stage and two-stage approaches.

One-stage REC methods: The pioneering work by Mao et al. (Mao et al. 2016) was the first work to focus on REC. Their method, an one-stage approach, utilizes CNN and LSTM to extract feature maps and text features from the image and expression, respectively. These features are then concatenated and input into an object detection model for region searching. Building upon this framework, subsequent works introduced various modules aimed at better aligning the expression and image. Recently, the transformer architecture has been applied to the REC task. Transformer-based REC methods (Deng et al. 2021; Ye et al. 2022; Kamath et al. 2021; Wang et al. 2022; Liu et al. 2023; Yang et al. 2023; Miao et al. 2024), particularly those employing one-stage designs, have achieved superior performance. However, in contrast to traditional one-stage methods, transformer-based methods rely on significantly higher memory usage and computational costs due to the self-attention mechanism. Furthermore, transformer-based approaches require large-scale datasets for effective model pre-training.

Two-stage REC methods: Diverging from one-stage methods, two-stage REC methods use a step-by-step process. Among two-stage REC methods, graph-based approaches (Yang, Li, and Yu 2019, 2020; Jing et al. 2020; Wang et al. 2019; Yang, Li, and Yu 2021; Wang et al. 2023a,b; Ke et al. 2024) consistently demonstrate superior performance as they consider not only the relationship between objects and expressions but also the relationships among objects themselves.

The methods mentioned above locate the object described by the expression based either on a set of predetermined learnable queries or the bounding box determined by the detector. When utilizing learnable queries, these approaches often encounter challenges in achieving convergence as they necessitate focusing on a specific region across the entire image. Additionally, these methods are less effective when handling small-sized target objects. On the other hand, when relying on bounding boxes determined by the detector, these methods fail to accurately localize the intended object if the detector does not precisely identify the object's location. Different from existing REC method, in this paper, we present a groundbreaking framework, called DiffusionREC for REC. In this method, we replace the learnable queries or detected bounding boxes with a set of random bounding boxes. Then, The generated bounding boxes are then incrementally denoised and refined by our filtering-based object decoder, which is adaptively conditioned on the expression, the results from previous denoised step and generated

vision-language features from our vision-language encoder. So, compared to existing REC methods, our DiffusionREC not only operates independently of detector performance but also achieves higher efficiency than methods using learnable queries for prediction.

Diffusion Model in Discriminative Tasks

Diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019; Song, Meng, and Ermon 2021) are a class of deep generative models that have been widely applied in computer vision (Yuan et al. 2024; Song et al. 2024; Gu et al. 2022; Avrahami, Lischinski, and Fried 2022), natural language processing (Gong et al. 2023; He et al. 2023; Reid, Hellendoorn, and Neubig 2023; Li et al. 2022), audio processing(Popov et al. 2021), and more. Additional applications of diffusion models can be found in recent surveys.

While diffusion models have achieved remarkable success in generative tasks, their potential in discriminative tasks is still emerging. Recent work has begun exploring diffusion models for segmentation (Baranchuk et al. 2022) and object detection (Chen et al. 2023), but their application to Referring Expression Comprehension (REC) remains largely unexplored. Despite advances in related fields like segmentation and object detection, progress in REC has been comparatively limited, with previous attempts to adapt generative diffusion models proving unsuccessful.

In this paper, we propose a novel diffusion-based approach for REC. Although our method shares some similarities with DiffusionDet, it incorporates several fundamental distinctions. DiffusionDet directly applies a diffusion model to object detection and utilizes a set prediction loss for bounding box denoising. In contrast, our approach incorporates an adaptive conditioning strategy to refine bounding box denoising at each step. To the best of our knowledge, this is the first work to employ a diffusion model for REC, highlighting its novelty and potential for improving comprehension in this domain.

Method

Fig 3 illustrates our REC framework, encompassing both the training and sampling processes. The denoising network for the proposed DiffusionREC consists primarily of two components: (1) an adaptive vision-text conditioning mechanism that conditions the noisy bounding box on the embedded visual features of object proposals from the previous denoising step, the embedded text features of the expression, and the vision-language features of the entire image and expression, and (2) a filtering-based object decoder that uses the interacted features to remove bounding boxes that do not cover the target object and adjust the bounding boxes that partially cover the target object during each denoising step. The DiffusionREC framework is elaborated as follows.

Diffusion Models

Diffusion models are a type of latent variable model, which utilize both a forward and a corresponding reverse diffusion process. Given an initial sample $x_0 \sim q(x_0)$, the forward process generates a Markov chain of latent variables

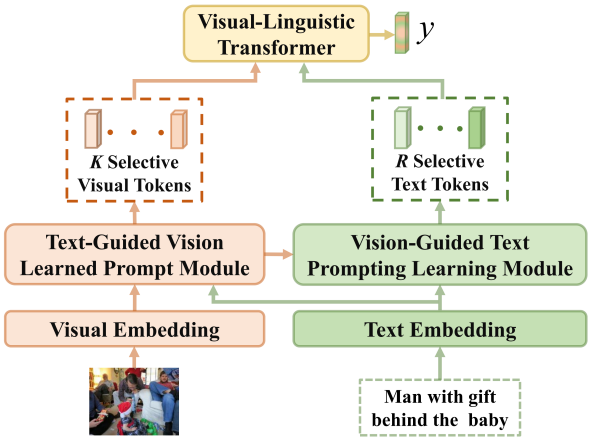

Figure 2: The process involves extracting visual features from objects and textual features from expressions. In contrast to existing visual-linguistic transformer encoder, our encoder incorporates a vision-guided text prompting learning module to selectively choose text features that convey crucial information during feature embedding.

$x_1, ..., x_T$ by iteratively adding small amounts of Gaussian noise to the sample:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right), \quad (1)$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ represents the variance schedule that controls the step size of the noise. After T steps, x_T approximates an isotropic Gaussian distribution. When β_t is sufficiently small, the reverse process $q(x_{t-1}|x_t)$ can be approximated by a Gaussian distribution, which can be learned using a parametric model:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (2)$$

where μ_θ and Σ_θ are trainable parameterized models. When conditioned on x_0, $q(x_{t-1}|x_t, x_0)$ has a closed-form solution. The corresponding objective function can be simplified to:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (3)$$

where $t \in \{1, ..., T\}$ denotes the time index of each denoising step, $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ represents Gaussian noise, ε_θ is the function for predicting noise from x_t, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$.

Vision-Guided Text Prompting Learning Module

Given an image-expression pair (I, r), the image I and expression r are first fed into the visual encoder and text encoder for embedding, respectively. The resulting embedded visual and text tokens are denoted as $[v_1, v_2, ..., v_{N_v}]$ and $[q_1, q_2, ..., q_{N_r}]$, respectively. Subsequently, these embedded visual and text tokens are respectively passed into the vision learned prompt module and our vision-guided text prompting learning module (VTPL) to identify and attend to tokens that carry important information. For the vision-learned prompt module, we follow the methodology described in (Rezaei et al. 2024; Zhou et al. 2023) to select

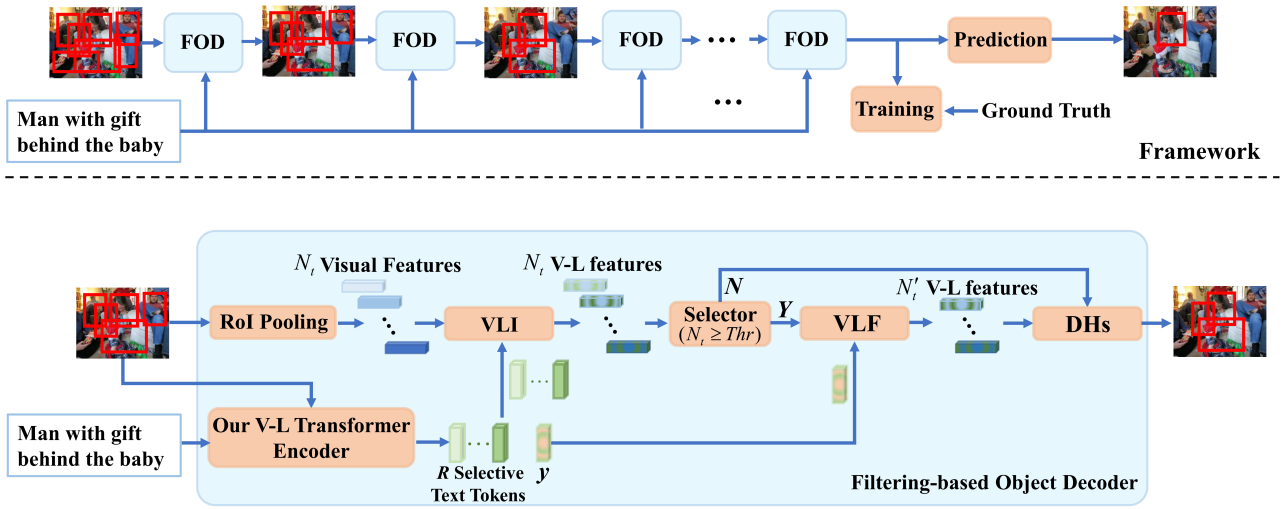

Figure 3: DiffusionREC uses a filtering-based object decoder (FOD) and a reverse process to refine bounding boxes. It treats boxes that partially cover or miss the target object as noise. In each reverse step, DiffusionREC corrects these inaccurate boxes and discards those covering only the background, ensuring only accurate boxes remain. **FOD** represents the filtering-based object decoder. **VLI** and **VLF** represents the vision-language interaction module and vision-language filtering module, respectively. **DHs** represents the detection heads. $N_t > N_t^{'}$. **V-L Transformer Encoder** represents the visual-linguistic transformer encoder. **V-L features** represents vision-language features.

K visual tokens that contain key information. In our VTPL approach, guided by the selected K visual tokens, each text token is paired with these visual tokens to compute a relevance score for the text token. Specifically, for the i-th text token, the corresponding score is calculated as follows.

$$\begin{aligned} [v_{l_1}, v_{l_2}, ..., v_{l_K}] &= \text{TVLP}\left([v_1, v_2, ..., v_{N_v}]\right), \\ \beta_{ij} &= \mathbf{W}_\beta \left[\tanh\left(\mathbf{W}_r q_i + \mathbf{W}_\ell v_{l_j}\right)\right], \\ \alpha_i &= \frac{\sum_{j=1}^{K} \beta_{ij}}{K}, \end{aligned} \tag{4}$$

where $\mathbf{W}_\beta$, $\mathbf{W}_r$ and $\mathbf{W}_\ell$ are trainable parameter matrices. TVLP represents the text-guided vision learned prompt module and α_i is the relevance score between the i-th text token and K selected visual token. Then, we sort the scores $\{\alpha_1, \alpha_2, ..., \alpha_{N_r}\}$ in descending order and select the text tokens with the top R scores for further processing. The text tokens that have been selected are represented as $e_R = [q_{n_1}, q_{n_2}, ..., q_{n_R}]$. Finally, these selected text tokens $[q_{n_1}, q_{n_2}, ..., q_{n_R}]$ are combined with the processed visual tokens $[v_{l_1}, v_{l_2}, ..., v_{l_K}]$ and inputted into the visual-linguistic transformer (e.g, Swin-Transformer, etc) for interaction. The output of the visual-linguistic transformer is denoted as y. The framework of our visual-linguistic transformer encoder is illustrated in Fig 2.

Forward and Reverse Processes for DiffusionREC

Forward Process. The framework of diffusionREC is showed in Fig 3. In our scenario, the data sample consists of N normalized ground truth bounding boxes, denoted as z_0. Specifically, z_0 represents the concatenation of N normalized ground truth bounding boxes (i.e., b_{GT}), where $\left(z_0 \in \mathbb{R}^{N\times 4}\right)$. In the t-th ($t > 0$) forward step, the noisy bounding boxes $\mathbf{z}_t$ are generated by introducing noise to $\mathbf{z}_{t-1}$ following the distribution $q\left(\mathbf{z}_t|\mathbf{z}_{t-1}\right) = \mathcal{N}\left(\mathbf{z}_t, \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}\right)$, where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ are the noise schedule. After completing T iterative steps, z_T is appropriately fine-tuned to ensure comprehensive coverage of the entire image.

Reverse Process. The reverse process involves generated the bounding boxes can precisely locate the target object through iterative refining and removing of the noisy bounding boxes from $\mathbf{z}_T \sim \mathcal{N}\left(0, \mathbf{I}\right)$. Let the text features of the expression and the vision-language features from the visual-linguistic transformer serve as the condition $\mathbf{c} = [e_R, y, \mathbf{v}_p]$, where $\mathbf{v}_p$ are the visual features of bounding boxes from previous denoised step. The reverse process can be modeled as a Markov chain with its joint distribution formulated as $p_\theta\left(\mathbf{z}_{0:T}|\mathbf{c}\right) := p\left(\mathbf{z}_T\right)\prod_{t=1}^{T} p_\theta\left(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}\right)$. Here, $\mathbf{z}_{t-1}$ is sampled from $p_\theta\left(\mathbf{z}_{0:T}|\mathbf{c}\right) := p\left(\mathbf{z}_T\right)\prod_{t=1}^{T} p_\theta\left(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}\right)$. p_θ is our filtering-based object decoder that θ are the parameters. Note that $\mathbf{c}$ is an adaptive condition because the vision-language features from the our vision-language encoder and the number of bounding boxes change at each reverse step. The details of our filtering-based object decoder are shown below.

Filtering-based Object Decoder

Our filtering-based object decoder design takes inspiration from sparse-RCNN (Sun et al. 2023b). During each denoising step, the filtering-based object decoder not only refines the bounding boxes remaining in current step but also filters out the ones that do not accurately locate the objects in the image. Assuming that there are only N_t bounding boxes remaining in the t-th denoising step, we proceed as follows.

First, we employ ROI pooling to extract the visual fea-

tures $\left[f_1^t, f_2^t, ..., f_{N_t}^t\right]$ of these N_t bounding boxes in the image.

Then, guided by the text features of the expression, we construct multi-modal features by integrating R selective text tokens with the visual features. Specifically, we compute R scores between the i-th visual feature and the R selective text tokens to capture their relationship. The computation details between the i-th visual feature and the j-th selective text token are as follows:

$$\begin{aligned} \hat{\alpha}_{ij}^t &= \mathbf{W}_\alpha \left[\tanh\left(\mathbf{W}_f f_{N_i}^t + \mathbf{W}_R q_{n_j}\right)\right], \\ \alpha_{ij}^t &= \frac{\hat{\alpha}_{ij}^t}{\sum_{k=1}^R \hat{\alpha}_{ik}^t}, \end{aligned} \tag{5}$$

where $\mathbf{W}_\alpha$, $\mathbf{W}_f$ and $\mathbf{W}_R$ are trainable parameter matrices. These computed scores $\{\alpha_{ik}^t\}_{k=1}^R$, R selective text tokens and the i-th visual feature are combined to form the i-th multi-modal feature m_i^t, where.

$$\begin{aligned} c_i^t &= \textstyle\sum_{k=1}^R \alpha_{ik}^t q_{n_k}, \\ m_i^t &= \mathbf{W}_{fc}\left[f_i, c_i^t\right]. \end{aligned} \tag{6}$$

Here, $\mathbf{W}_{fc}$ is a trainable parameter matrix. If N_t is smaller than the threshold T_r, the N_t multi-modal features $\{m_i^t\}_{i=1}^{N_t}$ are directly utilized for corresponding bounding boxes refinement by 8 detection heads with each head having 6 stages, following the setup of sparse R-CNN. Otherwise, We employ our newly introduced vision-language filtering module to mitigate the negative impact of multi-modal features that have low relevance to the vision-language feature from our vision-linguistic transformer encoder. Specifically, we compute the relationship scores between the vision-language feature y and the N_t multi-modal features $\{m_i^t\}_{i=1}^{N_t}$. The score between m_i^t and y is computed as

$$\gamma_i^t = \mathbf{W}_\gamma \left[\tanh\left(\mathbf{W}_m m_i^t + \mathbf{W}_y y\right)\right], \tag{7}$$

where $\mathbf{W}_\gamma$, $\mathbf{W}_m$ and $\mathbf{W}_y$ are trainable matrices. To filtering out the multi-modal features (bounding boxes) unrelated to the vision-language feature, we sort the scores $\{\gamma_i^t\}_{i=1}^{N_t}$ in descending order and filter out the multi-modal features with scores lower than $\frac{\xi \sum_{i=1}^{N_t} \gamma_i^t}{N_t}$. The remaining multi-modal features are then used for corresponding bounding boxes refinement by using the detection heads.

Training and Inference

Training. Due to the variation in the number of predicted bounding boxes at each denoised step, the condition in our DiffusionREC model can undergo changes during training. As a result, we are unable to directly compute the loss as we would in a standard diffusion model. Suppose that there are N_T bounding boxes remaining after T denosing steps. The multi-modal features of these bounding boxes are represented as $\left[m_1^T, m_2^T, ..., m_{N_T}^T\right]$. To compute the loss, we first calculate the similarity scores between the N_T multi-modal features and entire expression q. Then, we combine the bounding box of the multi-modal feature with the highest score and the ground truth bounding box to construct the smooth $\mathcal{L}_1$ loss, $\mathbf{SmoothL}_1(\cdot,\cdot)$. Therefore, the loss for our DiffusionREC is defined as follows:

$$\begin{aligned} \eta_i &= \left\langle \frac{\mathbf{W}_T m_{N_i}^T}{||\mathbf{W}_T m_{N_i}^T||}, \frac{\mathbf{W}_q q}{||\mathbf{W}_q q||} \right\rangle, \\ \pi_j &= \underset{j}{\operatorname{argmax}} \{\eta_j | 1 \le j \le N_T\}, \\ \mathcal{L}_{\text{reg}} &= \mathbf{SmoothL}_1(b_{\pi_j}, b_{GT}), \\ \mathcal{L} &= \mathcal{L}_{\text{reg}} + \lambda \frac{\eta_{\pi_j}}{\sum_{i=1}^{N_T} \eta_i}, \end{aligned} \tag{8}$$

where $\mathbf{W}_T$ and $\mathbf{W}_q$ are trainable parameter matrices. b_{π_j} is the bounding box of the π_j-th multi-modal feature and λ is the hyperparameter.

Inference. To begin, we randomly generate N bounding boxes that can fully cover the image. Subsequently, these bounding boxes undergo a gradual denoising process guided by the expression, vision-language features from our visual-linguistic transformer encoder, and the visual features of bounding boxes from the previous denoised step. After T denoising steps, the predicted bounding box is determined by selecting the one with the highest score, based on the similarity between the multi-modal features and the text features of the expression. If the Intersection-over-Union (IoU) value between the predicted bounding box and the ground-truth bounding box is larger than 0.5 , we consider the test sample to be predicted correctly.

Experimental Results

In this section, we provide extensive evaluation results for our proposed method on five challenging REC benchmarks: RefCOCO (Kazemzadeh et al. 2014), RefCOCO+ (Kazemzadeh et al. 2014), RefCOCOg (Mao et al. 2016), Flickr30K entities (Plummer et al. 2015) and ReferItGame (Kazemzadeh et al. 2014). RefCOCO and RefCOCO+ have expressions with average lengths of 3.61 and 3.65 words, respectively. On the other hand, RefCOCOg contains longer expressions, with averages of 8.4. Furthermore, the proportion of short expressions (i.e., containing only one or two noun chunks) in RefCOCO, RefCOCO+, RefCOCOg, and Ref-reasoning is 85.26%, 87.73%, 46.77%, and 19.68%, respectively. Due to limited space, we provide detailed information about these datasets in the supplementary materials.

Implementation Details

In this study, we adopt the same approach as described in (Deng et al. 2021) for embedding the image and expression. We set the number of randomly generated bounding boxes, denoted as N, to 1,000. Furthermore, during the training phase, we set the number of epochs, batch size, λ and learning rate to 50, 8, 0.5 and 10^{-4}, respectively. The denoised strategy we use is Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel 2020) with T=1,000 diffusion steps, and a square-root noise schedule. During evaluation, we found that the method performs best when the number of preserved bounding boxes is less than 3% of N ($Thr = 0.03N$). After T steps of denoising, around 30 bounding boxes are preserved.

Methods	Pre-train Dataset	RefCOCO			RefCOCO+			RefCOCOg		Flickr30k	Refclef
		val	testA	testB	val	testA	testB	val	test	test	test
M-DGT (Chen and Li 2022)	None	85.37	83.01	85.24	70.02	72.26	68.92	79.21	79.06	79.97	-
QRNet (Ye et al. 2022)	COCO	84.01	85.85	82.34	72.94	76.17	63.81	73.03	72.52	81.95	74.61
TransVG++ (Deng et al. 2023)	COCO	86.28	88.37	80.97	75.39	80.45	66.28	76.18	76.30	81.49	74.70
MDETR (Kamath et al. 2021)	COCO, VG, F30K	86.57	89.58	81.41	79.52	84.09	70.62	81.64	80.89	83.80	-
OFA-base (Wang et al. 2022)	VG	88.48	90.67	83.30	81.39	87.15	**74.29**	82.29	82.31	-	-
VG-LAW (Su et al. 2023b)	COCO	86.62	89.32	83.16	76.37	81.04	67.50	76.90	76.96	-	77.22
SMCIM (Miao et al. 2024)	COCO	85.10	88.23	80.08	74.44	879.48	65.21	77.25	75.78	-	75.18
LGR-NET (Lu et al. 2024)	COCO	85.63	88.24	82.69	75.32	80.60	68.30	76.82	77.03	81.97	74.64
Refcrowd (Qiu et al. 2022)	None	-	84.50	-	-	75.02	-	-	-	-	-
CLIPREC (Ke et al. 2024)	MIMI	-	84.63	84.51	-	76.82	63.07	-	76.83	-	-
CyCo (Wang, Deng, and Jia 2024)	ImageNet	**89.47**	**91.87**	85.33	80.40	87.07	69.87	81.31	81.04	-	-
LADS (Su et al. 2023a)	None	87.80	91.23	84.03	79.65	84.86	71.97	82.67	81.96	-	**78.82**
DiffusionREC w/ DDIM	COCO	88.07	90.85	87.03	**83.35**	87.33	72.58	83.35	**83.52**	85.95	77.64
DiffusionREC	COCO	89.35	91.54	**87.28**	83.04	**87.81**	72.79	**83.94**	83.50	**86.67**	78.14

Table 1: The comparison between our method and existing approaches on the RefCOCO, RefCOCO+, RefCOCOg, Flickr30K, RefClef, and Ref-Reasoning datasets. Compared transformer-based methods are pretrained on various datasets, including COCO (Pont-Tuset and Van Gool 2015), visual genome (Krishna et al. 2017), Flickr30K entities (Plummer et al. 2015), and MIMIC-CXR (Johnson et al. 2019). The best and second-best results are indicated in bold and underline, respectively. VG, F30K, and MIMI represent visual genome, Flickr30K entities, and MIMIC-CXR, respectively.

Evaluation Results

We present the evaluation results of our DiffusionREC as well as other compared methods on the RefCOCO, RefCOCO+, RefCOCOg, Flickr30K and Refclef datasets. The detailed results can be found in Table 1.

The results in Table 1 clearly demonstrate that our method outperforms both transformer-based and non-transformer-based approaches in most cases. This performance superiority is particularly evident in the testB set of RefCOCO (which comprises more incomparable short expressions) and the test set of Flickr30K entities. In these scenarios, our method consistently outperforms SMICM, LGR-NET, CLIPREC and CyCo in most cases, which are transformer-based methods released in 2024. Furthermore, our approach exhibits significantly superior performance compared to two unified architectures for multiple visual-linguistic tasks: OFA-base and MDETR. Our method achieves a performance improvement of 0.86% over OFA-base and 2.97% over MDETR, respectively. Notably, our proposed method even surpasses OFA-Base by 1.46% on the challenging test set of RefCOCOg, which consists of more complex expressions. Despite slightly lower performance on the testB set of RefCOCO+, our model does not require pre-training on a large-scale dataset and is more compact than OFA-Base. This makes our approach a competitive alternative to both OFA-Base and MDETR. To improve efficiency, we replaced DDPM with DDIM and reduced the denoising steps to 50, we can see that our method with DDIM performs similarly to DDPM but has a much faster inference time (1.13s vs. 22.6s).

Among the methods compared, M-DGT (Chen and Li 2022) and Refcrowd (Qiu et al. 2022) share similarities with our approach. M-DGT focuses on candidate-object pruning as a post-processing refinement, specifically targeting the local graph layout after each message-passing step. Refcrowd utilizes subject and attribute labels of expressions to progressively locate objects within the feature map of images. However, it is important to note that Refcrowd focuses primarily on identifying the 'person' within an image, whereas M-DGT utilizes learnable queries for prediction. In contrast, our DiffusionREC method adopts a different strategy. We aim to overcome the negative impact of the detector and learnable queries in REC by directly processing randomly generated bounding boxes. This allows our method to better match relevant objects compared to M-DGT, as evidenced by the results in Table 1. These findings further highlight the effectiveness of our proposed method.

Ablation Studies

In this subsection, we conduct various ablation studies to analyze the effectiveness of different components within the proposed method. Due to space limitations, additional ablation studies are provided in the supplementary materials.

DiffusionREC with and without VLF. We start by evaluating the performance of our method by comparing the inclusion and exclusion of the vision-language filtering (**VLF**) module. The results are displayed in Table 2. It is evident from the table that the performance of Diffusion without the **VLF** module is noticeably inferior to that with the **VLF** module.

DiffusionREC model using different bounding boxes initialization strategies. Building upon the DiffusionREC without the **VLF** module, we examine and compare the performance of DiffusionREC under different strategies for bounding box initialization. In one strategy, multiple bounding boxes (N) encompassing the entire image are generated and utilized for denoising. In the alternative strategy, only a single bounding box covering the entire image is employed. The findings showcased in Table 3 indicate that, on average, DiffusionREC achieves a performance improvement of approximately 6.61% when employing N bounding boxes for initialization, in contrast to using only one bounding box.

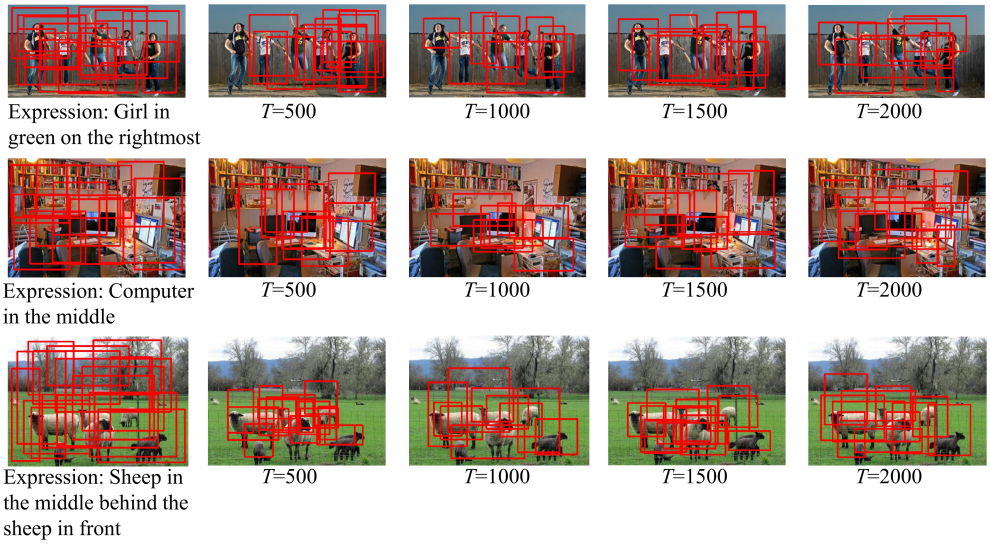

Figure 4: The visualizations of DiffusionREC at the 500th, 1,000th, 1,500th, and 2,000th denoised steps. The first and second rows illustrate successful predictions by DiffusionREC, showcasing corrected cases. The third row illustrates a failure case.

Methods	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
DiffusionREC w/o VLF	87.68	84.02	83.85	69.58	79.78
DiffusionREC w/ VLF	**91.54**	**87.28**	**87.81**	**72.79**	**83.50**

Table 2: Evaluation results of our DiffusionREC with and without **VLF** on RefCOCO, RefCOCO+, and RefCOCOg.

Methods	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
DiffusionREC w/ single bbx	83.37	80.03	82.72	65.18	77.81
DiffusionREC w/ N bbxes	**91.54**	**87.28**	**87.81**	**72.79**	**83.50**

Table 3: Evaluation results of our DiffusionREC model using different bounding box initialization strategies on RefCOCO, RefCOCO+, and RefCOCOg datasets.

DiffusionREC with varying initial bounding box settings. Next, assessing the performance of our method using different initial bounding box settings. The results in Table 4 indicate that when N is set to 100, 500, 1,000, and 1,500, the performance of DiffusionREC is slightly lower compared to when N is set to 1,000 in most cases.

Based on the outcomes of the aforementioned experiments, it becomes evident that the performance gains are not achieved by setting an excessive or insufficient number of bounding boxes during the initialization step. In the case of using too many bounding boxes, inaccuracies in their placement introduce substantial noise. Conversely, an insufficient number of bounding boxes fails to capture crucial information of the image effectively.

N				RefCOCO		RefCOCO+		RefCOCOg
100	500	1000	1500	testA	testB	testA	testB	test
✓				88.63	85.39	85.14	70.84	80.38
	✓			90.33	87.12	**87.92**	71.93	**83.54**
		✓		**91.54**	**87.28**	87.81	**72.79**	83.50
			✓	88.08	85.64	86.65	71.85	82.41

Table 4: The ablation results of DiffusionREC with varying initial bounding box settings N.

Visualization

In addition to the quantitative results, we also provide visual representations of the proposed method in Fig 4. The first and second rows depict the results of two scenarios of DiffusionREC at 500, 1,000, 1,500, and 2,000 denoised steps, while the third row showcases an error case. In the first row and second row, it can be observed that at the 500th denoised step, some of the bounding boxes that were only locating the background of the image have been removed. Meanwhile, the remaining bounding boxes have been refined to partially identify the target objects. As we progress to the 1,000th denoised step, more bounding boxes are further refined and accurately locate the target object. However, in the 1,500th and 2,000th steps, excessive fine-tuning leads to a shift in some bounding boxes. In the third row, despite the gradual refinement and movement of some bounding boxes towards the target object as the number of steps increases, they are unable to precisely locate the target object due to its similarity to neighboring objects.The aforementioned analyses shows that DiffusionREC effectively removes background-only bounding boxes and accurately locates the target object described by the expression during the denoising process.

Conclusion

In this paper, we present a groundbreaking framework, DiffusionREC, to address challenges in current REC methods, especially related to detector accuracy and learnable queries. Instead of discarding inaccurate bounding boxes, we treat them as noisy boxes. Through iterative denoising, we introduce a filtering-based detector that removes inaccurate boxes and refines the remaining ones. Additionally, we propose a novel visual-linguistic transformer encoder for enhanced integration of image and text features. Experiment results show that DiffusionREC performs better than current transformer-based REC methods on five datasets in most cases. With these notable advantages, DiffusionREC proves to be a powerful framework for the REC task.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-Driven Editing of Natural Images. In *CVPR*.

Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2022. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*.

Chen, S.; and Li, B. 2022. Multi-Modal Dynamic Graph Transformer for Visual Grounding. In *CVPR*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. DiffusionDet: Diffusion Model for Object Detection. In *ICCV*.

Dancette, C.; Whitehead, S.; Maheshwary, R.; Vedantam, R.; Scherer, S.; Chen, X.; Cord, M.; and Rohrbach, M. 2023. Improving Selective Visual Question Answering by Learning From Your Peers. In *CVPR*.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding With Transformers. In *ICCV*.

Deng, J.; Yang, Z.; Liu, D.; Chen, T.; Zhou, W.; Zhang, Y.; Li, H.; and Ouyang, W. 2023. TransVG++: End-to-End Visual Grounding With Language Conditioned Vision Transformer. *TPAMI*.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2024a. Hierarchical Local-Global Transformer for Temporal Sentence Grounding. *TMM*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2024b. Rethinking Weakly-supervised Video Temporal Grounding From a Game Perspective. In *ECCV*.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2023. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *ICLR*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *CVPR*.

Hamilton, M.; Zisserman, A.; Hershey, J. R.; and Freeman, W. T. 2024. Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Sound and Language. In *CVPR*.

He, Z.; Sun, T.; Tang, Q.; Wang, K.; Huang, X.; and Qiu, X. 2023. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models. In *ACL*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NIPS*.

Jia-Mu Sun, T. W.; and Gao, L. 2024. Recent advances in implicit representation-based 3D shape generation. *Visual Intelligence*.

Jing, C.; Wu, Y.; Pei, M.; Hu, Y.; Jia, Y.; and Wu, Q. 2020. Visual-Semantic Graph Matching for Visual Grounding. In *MM*.

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. In *CoRR*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *ICCV*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

Ke, J.; Wang, J.; Chen, J.-C.; Jhuo, I.-H.; Lin, C.-W.; and Lin, Y.-Y. 2024. CLIPREC: Graph-Based Domain Adaptive Network for Zero-Shot Referring Expression Comprehension. *TMM*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Li, H.; Ren, Z.; Guo, Y.; You, J.; and You, X. 2024a. LSVC: A Lifelong Learning Approach for Stream-View Clustering. *TNNLS*.

Li, X.; Pan, Y.; Sun, Y.; Sun, Y.; Sun, Q.; Ren, Z.; and Tsang, I. W. 2024b. Scalable unpaired multi-view clustering with Bipartite Graph Matching. *Information Fusion*.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-LM Improves Controllable Text Generation. In *NIPS*.

Liang, C.; Wang, W.; Zhou, T.; Miao, J.; Luo, Y.; and Yang, Y. 2022. Local-Global Context Aware Transformer for Language-Guided Video Segmentation. In *CoRR*.

Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R. K.; Mahadevan, V.; and Manmatha, R. 2023. PolyFormer: Referring Image Segmentation As Sequential Polygon Generation. In *CVPR*.

Lu, M.; Li, R.; Feng, F.; Ma, Z.; and Wang, X. 2024. LGR-NET: Language Guided Reasoning Network for Referring Expression Comprehension. *TCSVT*.

Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Feng, J.; Chao, H.; and Mei, T. 2023. Semantic-Conditional Diffusion Networks for Image Captioning. In *CVPR*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*.

Miao, P.; Su, W.; Wang, G.; Li, X.; and Xi, L. 2024. Self-Paced Multi-Grained Cross-Modal Interaction Modeling for Referring Expression Comprehension. *TIP*.

Peng, S.; Genova, K.; Jiang, C. ".; Tagliasacchi, A.; Pollefeys, M.; and Funkhouser, T. 2023. OpenScene: 3D Scene Understanding With Open Vocabularies. In *CVPR*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *ICCV*.

Pont-Tuset, J.; and Van Gool, L. 2015. Boosting Object Proposals: From Pascal to COCO. In *ICCV*.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *ICML*.

Qiu, H.; Li, H.; Zhao, T.; Wang, L.; Wu, Q.; and Meng, F. 2022. RefCrowd: Grounding the Target in Crowd with Referring Expressions. In *MM*.

Reid, M.; Hellendoorn, V. J.; and Neubig, G. 2023. DiffusER: Diffusion via Edit-based Reconstruction. In *ICLR*.

Rezaei, R.; Sabet, M. J.; Gu, J.; Rueckert, D.; Torr, P.; and Khakzar, A. 2024. Learning Visual Prompts for Guiding the Attention of Vision Transformers. In *CoRR*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Song, W.; Zhang, X.; Li, S.; Gao, Y.; Hao, A.; Hou, X.; Chen, C.; Li, N.; and Qin, H. 2024. HOIAnimator:Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models. In *CVPR*.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *NIPS*.

Su, W.; Miao, P.; Dou, H.; Fu, Y.; and Li, X. 2023a. Referring expression comprehension using language adaptive inference. In *AAAI*.

Su, W.; Miao, P.; Dou, H.; Wang, G.; Qiao, L.; Li, Z.; and Li, X. 2023b. Language Adaptive Weight Generation for Multi-Task Visual Grounding. In *CVPR*.

Sun, M.; Suo, W.; Wang, P.; Zhang, Y.; and Wu, Q. 2023a. A Proposal-Free One-Stage Framework for Referring Expression Comprehension and Generation via Dense Cross-Attention. *TMM*.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Yuan, Z.; and Luo, P. 2023b. Sparse R-CNN: An End-to-End Framework for Object Detection. *TPAMI*.

Tianyang Xu, X.-F. Z.; and Wu, X.-J. 2023. Learning spatio-temporal discriminative model for affine subspace based visual object tracking. *Visual Intelligence*.

Wang, J.; Ke, J.; Shuai, H.-H.; Li, Y.-H.; and Cheng, W.-H. 2023a. Referring Expression Comprehension Via Enhanced Cross-Modal Graph Attention Networks. *TOMM*.

Wang, J.; Shuai, H.-H.; Li, Y.-H.; and Cheng, W.-H. 2023b. Language-guided Residual Graph Attention Network and Data Augmentation for Visual Grounding. *TOMM*.

Wang, N.; Deng, J.; and Jia, M. 2024. Cycle-Consistency Learning for Captioning and Grounding. In *AAAI*.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019. Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks. In *CVPR*.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic Graph Attention for Referring Expression Comprehension. In *ICCV*.

Yang, S.; Li, G.; and Yu, Y. 2020. Graph-Structured Referring Expression Reasoning in the Wild. In *CVPR*.

Yang, S.; Li, G.; and Yu, Y. 2021. Relationship-Embedded Representation Learning for Grounding Referring Expressions. *TPAMI*.

Yang, X.; Che, H.; and Leung, M.-F. 2025. Tensor-based unsupervised feature selection for error-robust handling of unbalanced incomplete multi-view data. *Information Fusion*.

Yang, Z.; Kafle, K.; Dernoncourt, F.; and Ordonez, V. 2023. Improving Visual Grounding by Encouraging Consistent Gradient-Based Explanations. In *CVPR*.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding. In *CVPR*.

Yuan, H.; Zhang, S.; Wang, X.; Wei, Y.; Feng, T.; Pan, Y.; Zhang, Y.; Liu, Z.; Albanie, S.; and Ni, D. 2024. InstructVideo: Instructing Video Diffusion Models with Human Feedback. In *CVPR*.

Zhang, Y.; Luo, H.; and Lei, Y. 2024. Towards CLIP-driven Language-free 3D Visual Grounding via 2D-3D Relational Enhancement and Consistency. In *CVPR*.

Zhou, D.; Hou, Q.; Yang, L.; Jin, X.; and Feng, J. 2023. Token Selection is a Simple Booster for Vision Transformers. *TPAMI*.