APL: Anchor-based Prompt Learning for One-stage Weakly Supervised Referring Expression Comprehension

Yaxin Luo[1], Jiayi Ji[2], Xiaofu Chen[1], Yuxin Zhang[2], Tianhe Ren[3], and Gen Luo[⋆4]

[1] Technical University of Denmark
[2] Xiamen University
[3] International Digital Economy Academy
[4] Shanghai Artificial Intelligence Laboratory

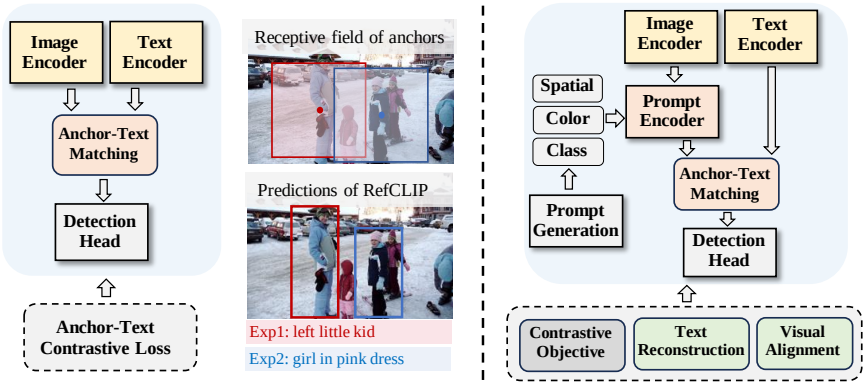
Abstract. Referring Expression Comprehension (REC) aims to ground the target object based on a given referring expression, which requires expensive instance-level annotations for training. To address this issue, recent advances explore an efficient one-stage weakly supervised REC model called RefCLIP. Particularly, RefCLIP utilizes anchor features of pre-trained one-stage detection networks to represent candidate objects and conducts anchor-text ranking to locate the referent. Despite the effectiveness, we identify that visual semantics of RefCLIP are ambiguous and insufficient for weakly supervised REC modeling. To address this issue, we propose a novel method that enriches visual semantics with various prompt information, called *anchor-based prompt learning* (APL). Specifically, APL contains an innovative *anchor-based prompt encoder* (APE) to produce discriminative prompts covering three aspects of REC modeling, *e.g.,* position, color and category. These prompts are dynamically fused into anchor features to improve the visual description power. In addition, we propose two novel auxiliary objectives to achieve accurate vision-language alignment in APL, namely text reconstruction loss and visual alignment loss. To validate APL, we conduct extensive experiments on four REC benchmarks, namely RefCOCO, RefCOCO+, RefCOCOg and ReferIt. Experimental results not only show the state-of-the-art performance of APL against existing methods on four benchmarks, *e.g.,* +6.44% over RefCLIP on RefCOCO, but also confirm its strong generalization ability on weakly supervised referring expression segmentation. Source codes released at: https://github.com/Yaxin9Luo/APL.

Keywords: Weakly Supervised Referring Expression Comprehension · Anchor-based Prompt Learning

1 Introduction

Referring Expression Comprehension (REC) aims to locate the target object based on a free-form language description [7, 29, 42]. As a fundamental vision-

⋆ Corresponding author



(a) Framework and visualization of RefCLIP (b) Framework of our APL

Fig. 1: Comparison of RefCLIP and APL. (a) RefCLIP adopts the anchor-text matching framework to conduct weakly supervised REC. However, anchor features often lead to the visual misleading in anchor-text matching. (b) APL overcomes the visual shortcoming of RefCLIP by fusing rich prompts into anchor features, and conduct anchor-based prompt learning via two auxiliary losses.

language task, REC has gained increasing attention and achieved significant progress recently [11, 29, 31, 38, 56, 64]. Driven by the great success, numerous efforts have been devoted to weakly supervised REC to reduce the expensive annotation costs [9, 33–35, 60, 62, 66]. Among them, most methods aim to directly apply weakly supervised objectives to a two-stage REC model, *e.g.,* MAttNet [64]. Despite the effectiveness, their two-stage modeling relies on region proposals and incurs expensive computational overhead.

Recently, Jin *et al.* [15] provided an efficient one-stage framework for weakly supervised REC, termed RefCLIP. As shown in Fig. 1, RefCLIP employs one-stage detectors to formulate weakly supervised REC as an anchor-text matching problem. Specifically, anchor (grid) features are first extracted from a pre-trained one-stage detector and then ranked based on the text features. To accomplish the REC task, RefCLIP will select the best matched anchor and decodes it to bounding box via the pre-trained detection head. During weakly supervised learning, RefCLIP adopts the anchor-text contrastive loss to achieve vision-language alignments using massive image-text pairs. Compared to two-stage methods, RefCLIP removes the expensive region proposal stage and achieves real-time inference speed, *e.g.,* +31.6 fps.

However, we identify that the visual semantics of RefCLIP are ambiguous and insufficient for weakly supervised REC modeling. Compared to region features, anchor features fall short in determining the actual visual object they represent. Despite the large receptive field of an anchor, it often encompasses noisy visual content and incomplete object information, which hinders the accurate anchor-text matching. As shown in Fig. 1, anchor features of the "*woman*" also capture the misleading information of the "*kid*", thereby matching with the incorrect expression "*left little kid*". Besides, we also notice that anchor features struggle to describe fine-grained visual semantics, *e.g.,* color, which are crucial for un-

derstanding diverse expressions. To explain, anchor features are pre-trained in common detection tasks, and the learned knowledge is fixed and limited, *e.g.,* 80 classes in COCO [27]. As shown in Fig. 1 (Exp2), RefCLIP fails to distinguish two kids that have fine-grained differences based on the expression.

To overcome these limitations, we propose a novel *anchor-based prompt learning* (APL) for one-stage weakly supervised REC. In particular, APL aims to improve the visual description ability of RefCLIP with an innovative *anchor-based prompt encoder* (APE). As shown in Fig. 1, APE can generate discriminative prompts that cover various knowledge of REC, *i.e.,* position, color and category. By injecting these prompts into anchor features, APE can greatly alleviate visual misleading in anchor-text matching. To effectively optimize APE, we further propose two novel auxiliary objectives in APL, namely text reconstruction loss and visual alignment loss. Specifically, text reconstruction loss minimizes the distance between text features and anchor features. And the visual alignment loss encourages anchor features to directly regress the pseudo-box[5] of the referent. With these training objectives, APL can achieve fine-grained vision-language alignments through weakly supervised training on massive image-text pairs.

To validate APL, we conduct extensive experiments on four common REC benchmarks, *i.e.,* RefCOCO [45], RefCOCO+ [45], RefCOCOg [44] and ReferItGame [17]. Experimental results show that APL achieves state-of-the-art performance on four datasets, *e.g.,* +6.44% over RefCLIP [15] on RefCOCO. Besides, we also conduct a bunch of ablation studies to validate our designs in APL. To validate the generalization ability of APL, we extend it to weakly supervised referring expression segmentation (RES) and also observe promising performance of APL against existing methods. In summary, our contributions are three folds:

- We identify the visual shortcoming in existing one-stage weakly supervised REC. To address this issue, we propose a novel *anchor-based prompt learning* (APL) to improve the visual description ability for the popular one-stage weakly supervised REC model termed RefCLIP.
- We propose an innovative *anchor-based prompt encoder* (APE) in APL, which generates and fuses rich multimodal prompts into anchor features. To achieve effective anchor-based prompt learning, we further equip APL with two auxiliary objectives, namely text reconstruction loss and visual alignment loss.
- APL achieves state-of-the-art results on four weakly supervised REC benchmark datasets. In addition, APL can be directly applied to weakly supervised RES and outperforms existing methods on three benchmark datasets.

2 Related Work

2.1 Referring Expression Comprehension

Referring Expression Comprehension (REC) aims to locate the target instance based on the given expression. Early REC methods [11,29,31,56,64] mainly follow

[5] The pseudo-box is produced via the anchor-text matching.

the two-stage pipeline, which first generates candidate regions using detection networks such as Faster-RCNN [54] and then selects the target one that best matches the referring expression. In spite of their success, two-stage methods are often criticized for their slow inference speed, which greatly limits their applications. To overcome this limitation, researchers have shifted their attention to one-stage REC [25, 39, 42, 70, 71]. In particular, these methods often embed text features into a one-stage detection network like YOLOv3 [53]. By directly predicting the target box via the detection head, one-stage REC models can achieve real-time inference speed. Based on this paradigm, existing methods further improve the reasoning ability via deep fusions [70] or attentions [39, 42, 71]. Subsequently, driven by the progress of Transformers [18, 40, 61], recent endeavors [7, 36, 39, 43] also explore their applications to one-stage REC, which often stacks multiple Transformer layers for better cross-modal interactions. In this paper, we explore weakly supervised learning for one-stage REC from a novel perspective of anchor-based prompt learning.

2.2 Weakly Supervised Referring Expression Comprehension

Compared to fully supervised REC, weakly supervised REC limits the access to ground-truth annotations. Most existing methods [9, 33–35, 60, 62, 66] often explore weakly training objectives to optimize traditional two-stage REC models with image-text pairs. Among them, sentence reconstruction [34, 35, 62] selects the best matched region to reconstruct the given input expression. Contrastive learning [9, 66] constructs positive and negative pairs from a set of regions and expressions and computes the InfoNCE loss [47]. Despite their effectiveness, these two-stage methods also suffer from expensive computational overhead. Therefore, researchers attempt to explore weakly supervised learning for one-stage REC [15, 67]. Among them, the advanced method called RefCLIP [15] adopts anchor-text matching based on the one-stage detector. To achieve weakly supervised learning, RefCLIP conducts anchor-based contrastive learning with massive image-text pairs.

In this paper, we identify that anchor representations often contain ambiguous object information and greatly hinder weakly supervised learning. To address this issue, we propose a novel *anchor-based prompt learning* (APL) for one-stage weakly supervised REC. APL aims to fuse rich prompt information into anchors and prompts vision-language alignments via two novel auxiliary objectives, *i.e.,* text reconstruction loss and visual alignment loss.

2.3 Prompt Learning

Prompt learning is an emerging research hot topic in natural language processing (NLP), which inserts text instructions into the input of a pre-trained language model for a better understanding of the task [14, 48, 58]. Early works [4, 6, 14, 48, 51] focus on manually selected prompts to improve the zero-shot and few-shot performance of language models. Recently, most works regard learnable

vectors as prompts and optimize them via task-specific fine-tuning. These methods can greatly improve the adaptation ability of language models to various downstream tasks [10, 12, 24, 37, 68, 69]. Inspired by these progresses, prompt learning has been a popular transfer learning scheme for pre-trained vision models. For example, VPT [68] adopts the deep prompt tuning strategy to transfer ViTs [8] to downstream tasks efficiently. CoOp [68] significantly improves the generalization ability of CLIP [50] on various out-domain tasks.

Different from previous work, APL dynamically constructs rich multimodal prompts, *e.g.,* position and color, to improve the anchor representations. We also introduce two novel objectives to achieve accurate anchor-based prompt learning.

3 Preliminary

We first recap the framework of RefCLIP [15], which defines weakly supervised REC as an anchor-text matching problem. In particular, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, anchor features $F_a \in \mathbb{R}^{(h\times w)\times d}$ are extracted from the last convolution feature map in YOLOv3 [53]. Based on anchor features, YOLOv3 employs the detection head to predict their corresponding bounding boxes. To accomplish REC, RefCLIP selects the target anchor that best matches with the given expression $T \in \mathbb{R}^L$, and predicts the bounding box of the referent via the detection head. This process can be formulated by

$$b = \mathcal{F}_{\text{det}}(\arg\max_{f_a \in F_a} \phi(f_a, f_t)), \tag{1}$$

where $f_a \in \mathbb{R}^d$ and $f_t \in \mathbb{R}^d$ denote anchor features and expression features, respectively. ϕ denotes the dot product similarity. And $\mathcal{F}_{\text{det}}(\cdot)$ is the detection head of YOLOv3 [53]. As defined in Eq. 1, once the target anchor is correctly selected, RefCLIP can directly predict the bounding box of the referent. Compared to two-stage methods, RefCLIP is much more efficient due to the elimination of the region proposal stage.

To achieve weakly supervised training, RefCLIP adopts anchor-text contrastive learning, defined by

$$\mathcal{L}_{\text{atc}} = -\log \frac{\exp\left(\phi(\hat{f}_{ai}, f_{ti})/\tau\right)}{\sum_{j=0}^{N} \mathbb{I}_{(i\neq j)} \exp\left(\phi(f_{aj}, f_{ti}))/\tau\right)}, \tag{2}$$

where $\hat{f}_{ai}$ denotes the best matched anchor features in i-th image, and N is the batch size. τ is the temperature for contrastive learning. With Eq. 2, RefCLIP can be directly optimized with massive image-text pairs.

As shown in Eq. 1, the effectiveness of RefCLIP lies in the accurate anchor-text matching. Nevertheless, anchor feature f_a suffers from the shortcoming of object representation. Compared to instance-level region features, anchor features are more fragmented and noisy, where an anchor often contains incomplete

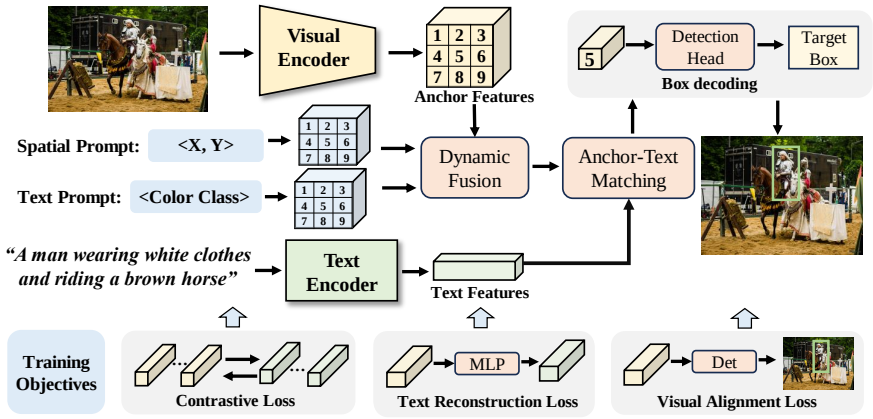


Fig. 2: Illustration of the proposed anchor-based prompt learning (APL). APL contains a novel anchor-based prompt encoder (APE), which fuses rich prompts into anchor features for accurate anchor-text matching. To promote anchor-based prompt learning, APL is equipped with two auxiliary losses, namely text reconstruction loss and visual alignment loss.

object information. Besides, anchor features also lack sufficient visual semantics for REC modeling, *e.g.,* color. Therefore, the visual shortcoming of anchor features inevitably hinders the vision-language alignment of RefCLIP.

4 The APL Framework

4.1 Overview

To address the above issues, we propose a novel weakly supervised REC framework, namely *anchor-based prompt learning* (APL). The core idea of APL is to improve object representations of RefCLIP [15] with various prompts. To achieve this target, APL is equipped with a novel *anchor-based prompt encoder* (APE) to generate discriminative prompts covering position, color and category. Then, these prompts are dynamically fused into anchor features. Therefore, the anchor-text matching process can be re-written by

$$b = \mathcal{F}_{\text{dec}}(\underset{f_a \in F_a, f_p \in F_p}{\arg\max}\ \phi(\mathcal{F}_{\text{prompt}}(f_a, f_p), f_t)). \tag{3}$$

Here, $\mathcal{F}_{\text{prompt}}(\cdot)$ denotes the anchor-based prompt encoder. $f_p \in \mathbb{R}^{h\times w\times d}$ is prompt features that describe position, attribute and category for anchor features $f_a \in \mathbb{R}^{h\times w\times d}$. Similar to RefCLIP, APL conducts the anchor-text ranking and the box decoding to locate the referent. For weakly supervised training, in addition to the contrastive objective of RefCLIP, we further propose two novel objectives to promote vision-language alignment in APL, *i.e.,* text reconstruction loss and visual alignment loss.

4.2 Anchor-based Prompt Encoder

As discussed above, anchor features often fall short of encoding discriminative object information. To tackle this challenge, the anchor-based prompt encoder (APE) first generates a set of prompt information for each candidate anchor. Then, these prompts are dynamically fused into anchor features to improve the visual description power.

Anchor-based prompt generation. The key to APE is how to generate valuable and discriminative prompts for different anchors. As shown in Fig. 2, we first define a prompt template that contains three slots, *i.e.,* position, color and category. Then, we extract these pieces of information from the image to obtain the anchor-based prompt.

In particular, given an anchor feature $f_a \in \mathbb{R}^d$, we obtain its corresponding detected box $b \in \mathbb{R}^4$ and category $c \in \mathbb{R}^1$ via the detection head. Then, the spatial prompt $f_{ps} \in \mathbb{R}^d$ is defined by

$$f_{ps} = \rho_{\text{pos}}(\frac{b_0 + b_2}{2}, \frac{b_1 + b_3}{2}), \tag{4}$$

where $\rho_{\text{pos}}(\cdot)$ is the positional embedding function [61]. f_{ps} indicates the spatial information for anchor a. As defined in Eq. 4, we first obtain the center coordinates of the detected bxo b and then transform it to spatial prompt via the positional embedding. In practice, the spatial prompt can also be represented as a word, *e.g.,* "left" and "right", but it can only reflect the rough location.

Afterward, we define the color and category prompts as natural language descriptions, *e.g.,* "black sofa" and "white cup". Specifically, the category can be directly obtained through the class name of c. The color information is calculated based on the pixel values of the image region b[6]. Then, we transform the obtained RGB color to natural words via a pre-defined color table. Finally, the color and the category are combined and processed to obtain textual features $f_{pt} \in \mathbb{R}^d$, which is defined by

$$f_{pt} = \mathcal{F}_{\text{text}}(t_p W_p), \tag{5}$$

where $t_p \in \mathbb{R}^{l_p}$ and $W_p \in \mathbb{R}^{l_p \times d}$ are tokenized prompt words and weights of the word embedding, respectively. As defined in Eq. 4 and 5, the generated prompts contain detailed object information, which can be combined with anchor features for better visual understanding.

Anchor-prompt fusion. To achieve the above target, we propose a dynamic weighting strategy to fuse the prompt features, *i.e.,* f_{ps} and f_{pt}, into the anchor features f_a, which is formulated by

$$f_p = w_0 f_a + w_1 f_{ps} + w_2 f_{pt}, \tag{6}$$

where $f_p \in \mathbb{R}^d$ is the anchor-based prompt features that are used to conduct anchor-text matching. $w_0, w_1, w_2 \in \mathbb{R}^1$ are three attention weights, defined by

$$w_0, w_1, w_2 = \text{softmax}(\sigma((f_a + f_{ps} + f_{pt})W_1)W_2). \tag{7}$$

[6] We adopt average pooling and K-means to capture colors of objects accurately.

Here, $W_1 \in \mathbb{R}^{d\times d}$ and $W_2 \in \mathbb{R}^{d\times 2}$ are two projection weights. $\sigma(\cdot)$ denotes the activation function of ReLU [1]. According to Eq. 7, weights of different features are dynamically adjusted for different samples.

4.3 Anchor-based Prompt Learning

Similar to RefCLIP, we adopt the contrastive loss to achieve weakly supervised learning. Besides, we propose two novel objectives to further facilitate the anchor-based prompt learning, namely text reconstruction loss and visual alignment loss. Therefore, the weakly supervised learning of APL can be written by

$$\min_{\theta} \mathcal{L}_{\text{ptc}}(F_p, f_t; \theta) + \mathcal{L}_{\text{tr}}(F_p, f_t; \theta) + \mathcal{L}_{\text{va}}(F_p; \theta), \tag{8}$$

where $\mathcal{L}_{\text{ptc}}$ is the anchor-text contrastive loss defined in Eq. 2. $\mathcal{L}_{\text{tr}}$ and $\mathcal{L}_{\text{va}}$ denote the text reconstruction loss and the visual alignment loss, respectively. θ denotes the model parameters.

In Eq. 8, $\mathcal{L}_{\text{tr}}$ aims to reconstruct expression features f_t with the best matched anchor-based prompt features $\hat{f}_p$, which is defined by

$$\mathcal{L}_{\text{tr}} = (\mathcal{F}_{\text{mlp}}(\hat{f}_p) - f_t)^2. \tag{9}$$

Here, $\mathcal{F}_{\text{mlp}}$ is a multi-layer MLP to project $\hat{f}_p$ into a latent space. In practice, $\hat{f}_p$ is selected from F_p based on the similarity of $\phi(f_p, f_t)$. With Eq. 9, anchor-based prompt features can learn fine-grained visual knowledge from diverse expressions and achieve better vision-language alignments.

In addition, we adopt the visual alignment loss to directly reconstruct the pseudo bounding box of the referent b based on F_p and f_t, which is defined by

$$\mathcal{L}_{\text{va}} = \mathcal{L}_{\text{det}}(\mathcal{F}_{\text{rec}}(F_p, f_t), b). \tag{10}$$

Here, $\mathcal{F}_{\text{rec}}$ denotes the REC decoder of SimREC [41], which contains a multimodal fusion layer, a GARAN layer [70] and a detection head [53]. $\mathcal{L}_{\text{det}}$ is the detection losses [26], which consists of the IoU loss [55] and the confidence loss [52]. Note that the bounding box b is generated via Eq. 3, which is still in line with the definition of weakly supervised learning. Since the bounding box b is often noisy at the beginning of training, we apply $\mathcal{L}_{\text{va}}$ after a short training phase. Compared to the text reconstruction loss, the visual alignment loss encourages anchor-based prompt features to encode discriminative semantics so that the target box can be directly predicted in a regression manner.

4.4 Network Settings

Feature extraction. We deploy APL based on RefCLIP [15]. In particular, we use a bi-directional GRU layer [3] and a self-attention layer [61] to extract text features $f_t \in \mathbb{R}^{l\times d}$ from the expression T. For the visual backbone, we employ DarkNet-53 [53] to process the input image I and obtain the anchor

Table 1: Ablation study of APL on *val* set of RefCOCO and RefCOCO+.

APE	$\mathcal{L}_{\text{ptc}}$	$\mathcal{L}_{\text{tr}}$	$\mathcal{L}_{\text{va}}$	RefCOCO	RefCOCO+
-	✓	-	-	60.36	40.39
✓	✓	-	-	62.89	41.54
✓	✓	✓	-	63.01	42.31
✓	✓	✓	✓	**64.51**	**42.71**

features $F_a \in \mathbb{R}^{h \times w \times d}$. Following RefCLIP, we also adopt multi-scale fusion to fuse anchor features of different layers. Then, we filter most anchor features with low detection confidences. And the remaining anchor features are fused with prompt features to obtain the anchor-based prompt features F_p.

Training and inference. For weakly supervised training, we directly sum three losses and optimize the model via backpropagation. During inference, APL has two different ways to predict the bounding box of the referent. The first way is based on the anchor-text ranking, as defined in Eq. 3. The second way adopts the predictive branch using Eq. 10, which calculates the visual alignment loss. We adopt the second one for inference, which performs slightly better.

5 Experiments

5.1 Datasets and Metrics

RefCOCO [45] contains 142,210 referring expressions and 50,000 objects in 19,994 images from MSCOCO [27]. Expressions of RefCOCO mainly describe about absolute position. **RefCOCO+** [45] consists of 141,564 referring expressions for 49,856 bounding boxes in 19,992 MSCOCO images. Different from RefCOCO, RefCOCO+ contains more descriptions of relationships and attributes. **RefCOCOg** [44] has 104,560 referring expressions and 54,822 bounding boxes for 26,711 images, where its expressions are longer and more complex than that of RefCOCO and RefCOCO+. RefCOCOg includes two different splits, *i.e.,* umd split [45] and google split [44]. We use the Google split in our experiments. ReferItGame [17] includes 120,072 referring expressions for 99,220 bounding boxes of 19,997 images. ReferItGame includes background descriptions, making it more challenging than RefCOCO and RefCOCO+.

For REC task, we use IoU@0.5 as the metric. In particular, a prediction is considered correct when the intersection over union (IoU) between the prediction and the ground truth is larger than 0.5. For the RES task, we follow previous works [2, 21, 28, 30, 49, 57, 63] to use mIoU as the metric, which averages the IoU scores of all testing samples.

5.2 Implementation Details

Following RefCLIP [15], the image resolution is set to 416×416, and the text length is 15, 15 and 20 for RefCOCO, RefCOCO+ and RefCOCOg, respectively.

Table 2: Ablation study of the anchor-based prompt encoder (APE) on RefCOCO and RefCOCO+. "S-prompt" and "T-prompt" denote the spatial prompt and the text prompt, respectively. Our final choice is colored in grey.

Settings	Choices	RefCOCO			RefCOCO+		
		val	testA	testB	val	testA	testB
S-prompt	Text	57.88	56.34	57.98	41.17	42.07	39.46
	Pos. Func.	**64.51**	**61.91**	**63.57**	**42.71**	**42.84**	**39.80**
T-prompt	BLIP [22]	63.47	60.72	63.16	42.11	41.84	39.44
	Template	**64.51**	**61.91**	**63.57**	**42.71**	**42.84**	**39.80**
Fusion	Add	**64.69**	61.75	63.67	42.10	42.00	38.89
	Concat	64.23	61.53	**64.08**	39.59	39.49	38.54
	Dynamic Sum	64.51	**61.91**	63.57	**42.71**	**42.84**	**39.80**

Table 3: Comparison of different prompt learning methods on three REC datasets. For fair comparisons, RefCLIP is used as the structure for VPT and CooP.

Method	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val-g
VPT [13]	55.39	53.30	53.94	33.56	33.74	31.72	40.79
CooP [68]	60.33	59.11	59.63	37.63	36.87	37.84	49.38
APL (ours)	**64.51**	**61.91**	**63.57**	**42.70**	**42.84**	**39.80**	**50.22**

We use YOLOv3 [53] pre-trained on MSCOCO[7] [27] as the detection network. In APL, dimensions of prompt features, anchor features and text features are set to 512. For the text reconstruction objective, we adopt a three-layer MLP with a hidden size 512 for projecting anchor features. During weakly supervised training, we use Adam [19] as the optimizer. And the learning rate and the batch size are set to 1e-4 and 64, respectively. Training consists of 25 epochs, and the visual alignment loss is applied after 9,000 steps. The remaining settings are kept the same with RefCLIP.

5.3 Quantitative Results

Ablation Studies. We conduct extensive experiments to validate designs of APL in Tab. 1, 2 and 3. In particular, Tab. 1 shows the cumulative ablations of APL. From this table, the first observation is that all designs obviously contribute to the final performance. Specifically, APE provides the most apparent gains of all designs, *e.g.,* +2.53% on RefCOCO, suggesting the significance of discriminative anchor semantics. With the help of auxiliary losses, the performance of APL can be further boosted, *e.g.,* +1.62% on RefCOCO. Besides, We also notice that the text reconstruction loss yields more significant improvements on the dataset containing more diverse expressions, *i.e.,* RefCOCO+. In contrast, the visual alignment loss offers more benefits on RefCOCO, where visual understanding is the main challenge. These results extensively validate the effectiveness of the proposed auxiliary losses for vision-language alignments.

[7] Validation and testing images in REC task are removed.

Table 4: Comparison with state-of-the-art methods on four REC benchmark datasets. *GT proposals* means that official annotations of MSCOCO are used as candidates. *Pseudo Label* denotes that the student REC model is trained using pseudo-labels generated by a teacher model. For example, RefCLIP_SimREC means that RefCLIP and SimREC are the teacher and the student, respectively.

Method	RefCOCO			RefCOCO+			RefCOCOg	ReferIt	Inference
	val	testA	testB	val	testA	testB	val-g	test	speed
GT Proposals:									
VC [46]CVPR'18	-	33.29	30.13	-	34.60	31.58	30.26	-	-
ARN [33]ICCV'19	38.05	36.43	36.47	34.53	36.40	36.12	39.62	-	-
KPRN [34]MM'19	36.34	35.28	37.72	37.16	36.06	39.29	38.37	33.87	-
DTWREG [60]TPAMI'21	39.21	41.14	37.72	39.18	40.01	38.08	43.24	-	-
EARN [32]TPAMI'22	38.08	38.25	38.59	37.54	37.58	37.92	45.33	36.86	-
RefCLIP_MAttNet	69.31	67.23	71.27	43.01	44.80	41.09	51.31	-	-
APL_MAttNet(ours)	**74.24**	**73.29**	**76.39**	**48.59**	**53.02**	**44.04**	**57.08**	-	-
Det Proposals:									
VC [46]CVPR'18	-	32.68	27.22	-	34.68	28.10	29.65	14.50	-
KAC Net [5]CVPR'18	-	-	-	-	-	-	-	15.83	-
MATN [67]CVPR'18	-	-	-	-	-	-	-	13.61	-
ARN [33]ICCV'19	32.17	35.25	30.28	32.78	34.35	32.13	33.09	26.19	5.7fps
IGN [66]NeurIPS'20	34.78	37.64	32.59	34.29	36.91	33.56	34.92	-	-
DTWREG [60]TAPAMI'21	38.35	39.51	37.01	38.91	39.91	37.09	42.54	-	5.9fps
RelR [35]CVPR'21	-	-	-	-	-	-	-	37.68	-
NCE+Dist [62]CVPR'21	-	-	-	-	-	-	-	38.39	-
RefCLIP [15]CVPR'23	60.36	58.58	57.13	40.39	40.45	38.86	47.87	39.58	31.3fps
APL (ours)	**64.51**	**61.91**	**63.57**	**42.70**	**42.84**	**39.80**	**50.22**	**41.80**	26.7fps
Pesudo Labels:									
RefCLIP_SimREC [15]	62.57	62.70	61.22	39.13	40.81	36.59	45.68	42.33	54.8fps
RefCLIP_Transvg [15]	64.08	63.67	63.93	39.32	39.54	36.29	45.70	42.64	19.3fps
APL_SimREC (ours)	63.94	64.72	61.21	**42.11**	**44.85**	**38.31**	**48.35**	**45.22**	54.8fps
APL_Transvg (ours)	**64.86**	**64.89**	63.87	39.28	41.08	36.45	46.11	43.25	19.3fps

In Tab. 2, we further compare different designs for anchor-based prompt encoder (APE). For the choice of the spatial prompt, we observe that the text description, *e.g.,* "left", performs much worse than the positional function, *e.g.,* -6.63% on RefCOCO. In practice, the positional function can provide more accurate and continuous spatial information than the text description. Besides, we attempt to use BLIP [22] to generate region captions as the text prompt, but the performance declines. To explain, the generated captions often include noisy and unrelated descriptions, which potentially causes the visual misleading. In Tab. 2, we compare different strategies for anchor-prompt fusions, and our dynamic fusion still outperforms other methods.

In Tab. 3, we compare APL with existing prompt learning methods, *i.e.,* VPT [13] and CooP [68]. The first observation is that APL greatly outperforms the other two methods on three datasets, *e.g.,* up to +5.97% on RefCOCO+ *testA*. From Tab. 3, we also find that VPT performs worse on three datasets, which achieves 55.39% on RefCOCO. To explain, prompts of VPT are not conditioned on anchors, which may produce useless and harmful prompt information. In contrast, CooP can generate anchor-based prompts and demonstrate promis-

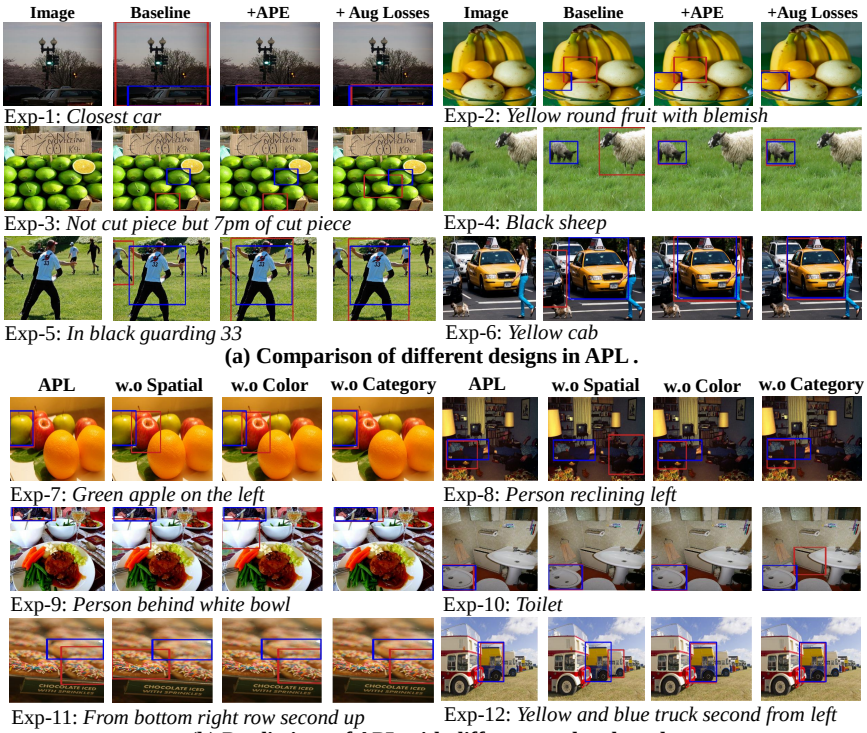
Table 5: Comparison of APL and existing methods on weakly supervised RES. PKS [21] uses click annotations for supervision, so we mark it in gray.

Method	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val-g
AMR [49]AAAI'22	14.12	11.69	17.47	14.13	11.47	18.13	15.83
GroupViT [63]CVPR'22	18.03	18.13	19.33	18.15	17.65	19.53	19.97
CLIP-ES [28]CVPR'23	13.79	15.23	12.87	14.57	16.01	13.53	14.16
GbS [2]ICCV'21	14.59	14.60	14.97	14.49	14.49	15.77	14.21
WWbL [57]NeurIPS'22	18.26	17.37	19.90	19.85	18.70	21.64	21.84
TSEG [59]arXiv'20	30.12	-	-	25.95	-	-	22.62
ALBEF [23]NeurIPS'21	23.11	22.79	23.42	22.44	22.07	22.51	24.18
I-Chunk [20]ICCV'23	31.06	32.30	30.11	31.28	32.11	30.13	32.88
TRIS [30]ICCV'23	31.17	32.43	29.56	30.90	30.42	30.80	36.00
PKS [21] arXiv'22	49.27	52.23	45.64	37.79	42.09	32.87	36.43
APL (ours)	**55.92**	**54.84**	**55.64**	**34.92**	**34.87**	**35.61**	**40.13**

ing performance on RefCOCOg. Nevertheless, CooP is limited in weak spatial and attribute information, so it still lags behind APL by large margins on RefCOCO and RefCOCO+. These results greatly validate the design of APL.

Comparison with existing methods on REC. In Tab. 4, we compare APL with a set of methods on four REC benchmark datasets. From this table, we find that most methods adopt the two-stage modeling and their inference speed is vastly inferior to the one-stage one, *e.g.,* 5.9 fps of DTWREG [60] *vs.* 26.7 fps of APL. Regarding performance, one-stage models have obvious advantages in RefCOCO, suggesting its better spatial understanding ability. Nevertheless, on more challenging datasets like RefCOCO+ and ReferIt, one-stage models perform similarly to the two-stage ones. As discussed in Sec.1, existing one-stage models often adopt a simple anchor-text matching, which lacks sufficient visual semantics for fine-grained REC modeling. Compared to these methods, APL achieves the best performances on four datasets, *e.g.,* +6.44% over RefCLIP on RefCOCO, and also maintains remarkable inference efficiency, *i.e.,* 26.7 fps. Compared to RefCLIP, APL has obvious performance gains on some complex splits, *e.g.,* RefCOCO *testB* and RefCOCOg *val*. Notably, APL can even achieve comparable performance with early supervised REC models on RefCOCO, *e.g.,* 63.57 of APL *vs.* 64.85 of Spe+Lis+Rl [65] on *testB*. These results further confirm that our APL can greatly promote fine-grained vision-language alignments.

In Tab. 4, we further validate APL under the pseudo-label learning setups. Following RefCLIP [15], we use the well-trained APL to generate pseudo-labels for training common REC models. As shown in Tab. 4, the REC model taught by APL greatly outperforms the one taught by RefCLIP, *e.g.,* up to +4.04% for SimREC [41]. Besides, we also notice that APL brings more benefits for the student REC model on challenging datasets like ReferIt and RefCOCOg. In particular, SimREC supervised by APL can even outperform APL by +3.42% in ReferIt, suggesting that APL can produce high-quality pseudo-labels in this dataset. Nevertheless, we also find that performance gains of TransVG are not obvious as SimREC. We conjecture that the Transformer structure of TransVG has higher requirements for data quality.



(a) Comparison of different designs in APL .

(b) Predictions of APL with different anchor-based prompts.

Fig. 3: Visualizations of APL. Subfig-(a) compares different designs of APL and visualizes their predictions. Subfig-(b) compares predictions of APL with different prompts. Predictions and ground-truths are colored in red and blue, respectively.

Generalization results on RES. In Tab. 5, we extend APL to weakly supervised RES and compare it with a set of existing methods. To accomplish the RES task, we use the YOLOv5-Seg [16] as the detection network in APL, which can produce segmentation masks based on anchor features. From Tab. 5, we observe that APL demonstrates obvious performance gains against existing methods. For example, APL outperforms the state-of-the-art method, *i.e.,* TRIS [30], by up to 26.08%, which is a remarkable improvement. In other RES datasets, the obvious benefits of APL can also be witnessed, *e.g.,* +4.81% and +4.13% in RefCOCO+ and RefCOCOg, respectively. Compared to PKS [21], which uses additional click annotations for supervision, APL still performs better in RefCOCO and RefCOCOg.

5.4 Qualitative Analysis

To gain in-depth insights into APL, we visualize and compare predictions of APL in Fig. 3. In Fig. 3 (a), we ablate the proposed APE and auxiliary losses and visualize the predictions. This figure shows that the default baseline often falls short in fine-grained recognition, *e.g.,* "*black sheep*" of Exp-4. With the help of

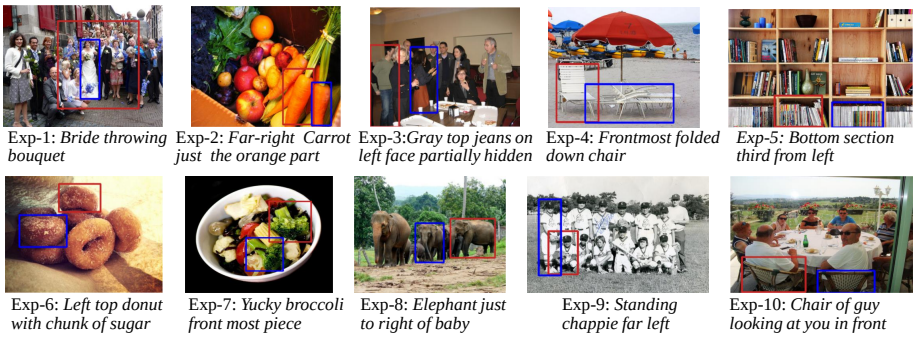


Fig. 4: Failure cases of APL on RefCOCO, RefCOCO+ and RefCOCOg.

APE, APL can better capture visual appearances such as color. However, APL still fails in some examples that involve complex descriptions, *e.g.,* Exp-2. These examples are correctly predicted after adopting two auxiliary losses. As shown in Fig. 3 (a) Exp-2, the "*orange*" can be located from a bunch of fruits.

In Fig. 3 (b), we further compare the effects of different prompt information in APE. As shown in Fig. 3 (b), removing the spatial information leads to the failure of APL on spatial-related expressions, *e.g.,* "*person reclining left*" of Exp-8. Besides, we observe that the color prompt is significant for locating the referent in a diverse scenario. Without the color prompt, APL can not correctly ground the "*green apple*" from various fruits in Exp-7. Moreover, the category prompt helps APL distinguish objects of similar appearances, *e.g.,* the "*toilet*".

To better understand the limitation of APL, we visualize its failure cases in Fig. 4. From this figure, we observe that APL still struggles to address long expressions, *e.g.,* Exp-3 and Exp-10, which requires strong reasoning ability. Besides, we can also see that complex visual scenes cause some failure cases, *e.g.,* the occluded objects in Exp-2. From these examples, we believe that APL has much room for improvement in reasoning ability and visual understanding.

6 Conclusion

In this paper, we focus on one-stage weakly supervised REC and identify that existing methods suffer from the ambiguous visual semantics in their REC modeling. To address this issue, we propose a novel approach, namely anchor-based prompt learning (APL). APL formulates weakly supervised REC as an anchor-text matching problem, equipped with an innovative anchor-based prompt encoder (APE) to enrich anchor semantics with a set of prompts. Moreover, we propose two novel auxiliary losses to achieve the effective anchor-based prompt learning, namely text reconstruction loss and visual alignment loss. Experimental results not only validate the state-of-the-art performance of APL in REC but also confirm the strong generalization ability of APL in RES.

Acknowledgements: This work was supported by the National Natural Science Foundation of China (No. 623B2088).

References

1. Agarap, A.F.M.: Deep learning using rectified linear units (relu) (2018) 8
2. Arbelle, A., Doveh, S., Alfassy, A., Shtok, J., Lev, G., Schwartz, E., Kuehne, H., Barak Levi, H., Sattigeri, P., Panda, R., Chen, C.F., Bronstein, A., Saenko, K., Ullman, S., Giryes, R., Feris, R., Karlinsky, L.: Detector-free weakly supervised grounding by separation (2021) 9, 12
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014) 8
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. pp. 1877–1901 (2020) 4
5. Chen, K., Gao, J., Nevatia, R.: Knowledge aided consistency for weakly supervised phrase grounding (2018) 11
6. Chen, Y., Liu, Y., Dong, L., Wang, S., Zhu, C., Zeng, M., Zhang, Y.: AdaPrompt: Adaptive model training for prompt-based NLP. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 6057–6068. Association for Computational Linguistics (2022) 4
7. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. pp. 1769–1779 (2021) 1, 4
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020) 5
9. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. pp. 752–768 (2020) 2, 4
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp (2019) 5
11. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. pp. 108–124 (2016) 2, 3
12. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models (2022) 5
13. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning (2022) 10, 11
14. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics (TACL) (2020) 4
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. pp. 2681–2690 (2023) 2, 3, 4, 5, 6, 8, 9, 11, 12
16. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., et al.: Ultralytics/yolov5: V7.0 - yolov5 sota realtime instance segmentation. Zenodo (2022) 13
17. 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) 3, 9

18. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. pp. 5583–5594 (2021) 4
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014) 10
20. Lee, J., Lee, S., Nam, J., Yu, S., Do, J., Taghavi, T.: Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21870–21881 (2023) 12
21. Li, H., Sun, M., Xiao, J., Lim, E.G., Zhao, Y.: Fully and weakly supervised referring expression segmentation with end-to-end learning (2022) 9, 12, 13
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022) 10, 11
23. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 9694–9705. Curran Associates, Inc. (2021) 12
24. Li, S.M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test time prompt tuning for zero-shot generalization in vision language models (2022) 5
25. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. pp. 10880–10889 (2020) 4
26. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 8
27. 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: European Conference on Computer Vision. pp. 740–755. Springer (2014) 3, 9, 10
28. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation (2022) 9, 12
29. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. pp. 4670–4679 (2019) 1, 2, 3
30. Liu, F., Liu, Y., Kong, Y., Xu, K., Zhang, L., Yin, B., Hancke, G., Lau, R.: Referring image segmentation using text supervision (2023) 9, 12, 13
31. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. pp. 1950–1959 (2019) 2, 3
32. Liu, X., Li, L., Wang, S., Zha, Z.J., Li, Z., Tian, Q., Huang, Q.: Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding (2022) 11
33. Liu, X., Li, L., Wang, S., Zha, Z.J., Meng, D., Huang, Q.: Adaptive reconstruction network for weakly supervised referring expression grounding. pp. 2611–2620 (2019) 2, 4, 11
34. Liu, X., Li, L., Wang, S., Zha, Z.J., Su, L., Huang, Q.: Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. pp. 539–547 (2019) 2, 4, 11
35. Liu, Y., Wan, B., Ma, L., He, X.: Relation-aware instance refinement for weakly supervised visual grounding. pp. 5612–5621 (2021) 2, 4, 11
36. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. pp. 13–23 (2019) 4

37. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5206–5215 (2022) 5
38. Luo, G., Zhou, Y., Ji, R.: Towards language-guided visual recognition via dynamic convolutions. International Journal of Computer Vision **132**(1), 1–19 (2024) 2
39. Luo, G., Zhou, Y., Ji, R., Sun, X., Su, J., Lin, C.W., Tian, Q.: Cascade grouped attention network for referring expression segmentation. pp. 1274–1282 (2020) 4
40. Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R.: Cheap and quick: Efficient vision-language instruction tuning for large language models. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (2023) 4
41. Luo, G., Zhou, Y., Sun, J., Huang, S., Sun, X., Ye, Q., Wu, Y., Ji, R.: What goes beyond multi-modal fusion in one-stage referring expression comprehension: An empirical study (2022) 8, 12
42. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. pp. 10034–10043 (2020) 1, 4
43. Luo, G., Zhou, Y., Sun, X., Wang, Y., Cao, L., Wu, Y., Huang, F., Ji, R.: Towards lightweight transformer via group-wise transformation for vision-and-language tasks. IEEE Transactions on Image Processing (2022) 4
44. 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) 3, 9
45. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: European Conference on Computer Vision. pp. 792–807. Springer (2016) 3, 9
46. 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 (2021) 11
47. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 4
48. Petroni, F., Rocktaschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S.: Language models as knowledge bases? (2019) 4
49. Qin, J., Wu, J., Xiao, X., Li, L., Wang, X.: Activation modulation and recalibration scheme for weakly supervised semantic segmentation (2021) 9, 12
50. Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 5
51. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research (JMLR) (2019) 4
52. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 8
53. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 4, 5, 8, 10
54. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 4

55. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 8
56. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. pp. 817–834 (2016) 2, 3
57. Shaharabany, T., Tewel, Y., Wolf, L.: What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs (2022) 9, 12
58. Shin, T., Razeghi, Y., IV, R.L.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. pp. 4222–4235 (2020) 4
59. Strudel, R., Laptev, I., Schmid, C.: Weakly-supervised segmentation of referring expressions (2022) 12
60. 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**(1), 4189–4195 (2021) 2, 4, 11, 12
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 4, 7, 8
62. Wang, L., Huang, J., Li, Y., Xu, K., Yang, Z., Yu, D.: Improving weakly supervised visual grounding by contrastive knowledge distillation. pp. 14090–14100 (2021) 2, 4, 11
63. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision (2022) 9, 12
64. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. pp. 1307–1315 (2018) 2, 3
65. Yu, L., Tan, H., Bansal, M., Berg, T.L.: A joint speaker-listener-reinforcer model for referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 12
66. 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) 2, 4, 11
67. Zhao, F., Li, J., Zhao, J., Feng, J.: Weakly supervised phrase localization with multi-scale anchored transformer network. pp. 5696–5705 (2018) 4, 11
68. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models (2021) 5, 10, 11
69. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022) 5
70. Zhou, Y., Ji, R., Luo, G., Sun, X., Su, J., Ding, X., Lin, C.W., Tian, Q.: A real-time global inference network for one-stage referring expression comprehension. IEEE Transactions on Neural Networks and Learning Systems (2021) 4, 8
71. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. arXiv preprint arXiv:2203.16265 (2022) 4