This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DViN: Dynamic Visual Routing Network for Weakly Supervised Referring Expression Comprehension

Xiaofu Chen[1] Yaxin Luo[1] Gen Luo[2*] Jiayi Ji[3] Henghui Ding[4] Yiyi Zhou[3]
[1]MBZUAI [2]OpenGVLab, Shanghai AI Laboratory [3]Xiamen University [4]Fudan University

Abstract

In this paper, we focus on weakly supervised referring expression comprehension (REC), and identify that the lack of fine-grained visual capability greatly limits the upper performance bound of existing methods. To address this issue, we propose a novel framework for weakly supervised REC, namely Dynamic Visual routing Network (DViN), which overcomes the visual shortcomings from the perspective of feature combination and alignment. In particular, DViN is equipped with a novel sparse routing mechanism to efficiently combine features of multiple visual encoders in a dynamic manner, thus improving the visual descriptive power. Besides, we further propose an innovative weakly supervised objective, namely Routing-based Feature Alignment (RFA), which facilitates the visual understanding of routed features through the intra-modal and inter-modal alignment. To validate DViN, we conduct extensive experiments on four REC benchmark datasets. Experiments demonstrate that DViN achieves state-of-the-art results on four benchmarks while maintaining competitive inference efficiency. Besides, the strong generalization ability of DViN is also validated on weakly supervised referring expression segmentation. Source codes are anonymously released at: https://github.com/XxFChen/DViN.

1. Introduction

Referring Expression Comprehension (REC), also known as visual grounding, is the task of locating target objects in images based on a given natural language description [4, 22]. As a vision-language (VL) task, REC has shown great potential in various real-world applications, *e.g.,* robot navigation [9] and image editing [33]. Therefore, numerous efforts have been devoted to REC recently. Despite the progress, most REC methods [8, 26, 52] usually require extensive object-level labels for their supervised learning, leading to expensive annotation costs.

To address this challenge, increasing attention has been focused on weakly supervised learning for REC [5, 10, 24, 25, 28, 47, 49]. Among them, early works [5, 24] often formulated weakly supervised REC using a two-stage pipeline, which requires an expensive region proposal stage. Recently, the one-stage framework [10, 54] has become a popular paradigm for weakly supervised REC due to its superior efficiency. For example, RefCLIP [10] utilizes anchor features from a pre-trained one-stage detector to represent candidate objects and directly matches them with text features to determine the target referent. Based on this efficient framework, contrastive-based objectives are used to achieve effective weakly supervised learning.

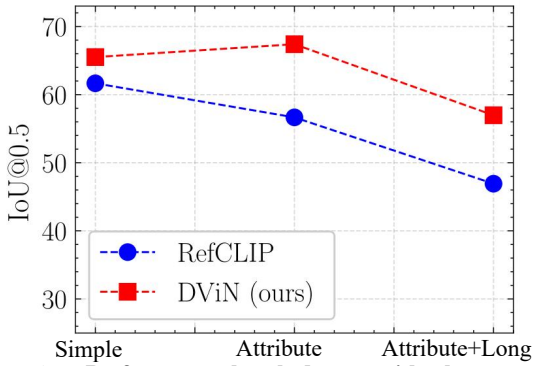
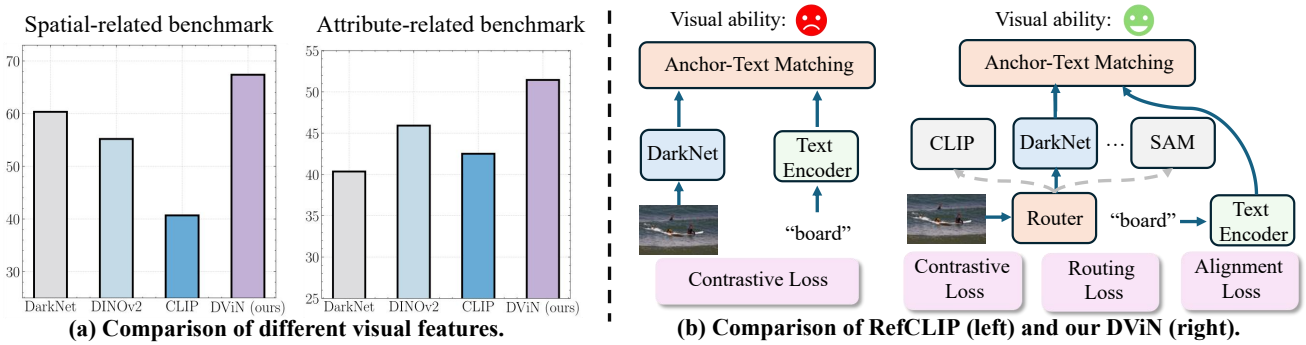


Figure 1. **Performance break-downs with the expression difficulty.** "Simple" refers to the group of short expressions (length≤20) without attributes. Compared to the "simple" group, expressions of "Attribute" group further contain attribute words. "Attribute+Long" group contains longer expressions (length>20) than "Attribute" group.

Despite the effectiveness, existing one-stage weakly supervised REC remains significantly inferior in fine-grained object recognition. As shown in Fig. 1, the state-of-the-art approach called RefCLIP achieves promising accuracy on simple descriptions, but greatly falls short on fine-grained ones, *e.g.,* attribute descriptions. To explain, visual features of existing methods are pre-trained on limited categories, *e.g.,* 80 classes of MS-COCO [20], limiting their fine-grained recognition ability. Moreover, without accurate annotations, weakly supervised approaches [18, 39] struggle with efficiently learning novel instance-level visual con-

*Corresponding author



(a) Comparison of different visual features. **(b) Comparison of RefCLIP (left) and our DViN (right).**

Figure 2. **Comparison of existing weakly supervised REC methods with our DViN.** (a) Visual features demonstrate orthogonal benefits on the spatial-related benchmark (RefCOCO) and the attribute-related benchmark (RefCOCO+). (b) Compared to RefCLIP, our DViN improves the visual descriptive ability via its dynamic visual routing, which can select the optimal visual features for anchor-text matching.

cepts from noisy image-text pairs.

To address this limitation, we propose a novel **D**ynamic **V**isual Routi**n**g **N**etwork (DViN) for weakly supervised REC. Our motivation is that combining different visual features can achieve complementary benefits in weakly supervised REC. As shown in Fig. 2, CLIP [39] and DINOv2 [37] pre-trained on various categories can capture fine-grained visual information, while DarkNet [40] shows promising spatial-aware ability. Motivated by this, DViN formulates visual feature extraction as a routing process, which dynamically combines visual encoders to maximize the visual understanding. To retain the efficiency, only few visual encoders are activated in one routing process. Meanwhile, we propose a novel objective to optimize the routing process, namely diversity-aware routing learning, which encourages the routing diversity and efficiency during training.

Based on DViN, we propose a novel weakly supervised objective to further improve the visual description power in the routing process, namely Routing-based Feature Alignment (RFA). Specifically, RFA includes a text-guided alignment loss and a vision-guided alignment loss, which enrich routed feature representation through the intra-modal and inter-modal alignment. As shown in Fig. 3, the text-guided alignment loss minimizes the distance between routed visual features and textual features, thus better capturing fine-grained semantics of the textual expression. Besides, the vision-guided alignment loss further optimizes the visual representation by encouraging a single visual feature to approximate the dynamic combination of visual features.

To validate DViN, we conduct comprehensive experiments on four REC benchmarks including RefCOCO [35], RefCOCO+ [35], RefCOCOg [34], and ReferItGame [12]. Experimental results not only demonstrate the significant performance gains of DViN on RefCLIP, *e.g.,* +16.53% on RefCOCO+, but also confirm its superior efficiency than a set of baselines, *e.g.,* +10.0 fps. More importantly, the strong generalization ability of DViN is also validated on weakly supervised referring expression segmentation (RES). In summary, our contributions are three-fold:

- We identify that the fine-grained recognition ability of existing weakly supervised REC models can be greatly improved via the mixture of visual features. Based on this finding, we propose a novel Dynamic Visual Routing Network (DViN).
- In DViN, we propose two innovative designs to maximize effectiveness while maintaining efficiency, namely the vision-conditioned router and diversity-aware routing learning. Additionally, we introduce the Routing-based Feature Alignment (RFA) to further enrich visual representations from the perspective of optimization.
- DViN achieves state-of-the-art results on four weakly supervised REC benchmark datasets. Besides, its generalization ability is validated on weakly supervised RES.

2. Related Work

2.1. Referring Expression Comprehension

Referring Expression Comprehension (REC) [6] refers to the task of locating a specific target in the image based on given natural language expression. There are two main methodologies for REC, *i.e.,* two-stage and one-stage methods. Two-stage methods [8, 22, 26, 42, 52] first generate a set of potential object regions using detection networks like Faster-RCNN [41] and then rank these regions to find the best match for the descriptive phrase. Recently, researchers have focused on one-stage methods due to their efficiency [19, 30, 31, 55, 56]. By predicting the target object in one step, these methods significantly improve processing speed. Early one-stage methods [30, 31, 55, 56] focused on shallow multimodal fusion, but recent advancements have leveraged the power of Transformer architectures [13, 48]. These newer models can enhance cross-modal interactions and improve performance [4, 29]. Meanwhile, some researchers have also explored using large language models for the REC task [46], leveraging their strong reasoning and multimodal understanding capabili-

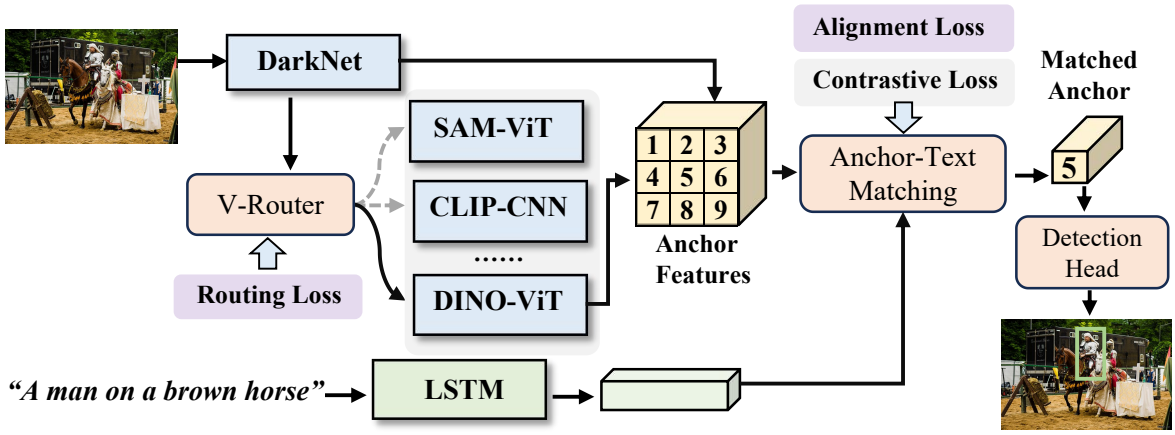


Figure 3. **Overview the proposed DViN framework.** In DViN, the vision-conditioned router will dynamically select the optimal visual encoder to refine the anchor features extracted by DarkNet. After that, anchor-text matching is conducted to locate the referent. DViN adopts the anchor-text contrastive loss for weakly supervised learning, which also includes an routing loss for efficient routing learning and an alignment loss for visual representation learning.

ties. However, both one-stage and two-stage methods still require expensive annotations for weakly supervised learning, limiting their real-world applications.

2.2. Weakly Supervised Referring Expression Comprehension

Unlike fully supervised methods, where every object in the image is tagged with its corresponding referring expression, weakly supervised REC aims to perform this task with text-only information. Most current approaches [10, 54]are inspired by two-stage supervised REC models and treat weakly supervised REC as a region-text ranking problem. Based on this framework, researchers have deployed various weakly supervised objectives, such as sentence reconstruction [26, 28, 49] and contrastive learning [5, 53]. Despite their effectiveness, two-stage methods often require significant computational resources. To overcome this, recent research has explored one-stage methods for weakly supervised REC. For example, RefCLIP [10] uses anchor-text matching within a one-stage detector framework, which employs anchor-based contrastive learning for supervision. However, recent weakly supervised REC methods are still limited by their visual representations and fall short in fine-grained recognition.

3. Preliminary

We first recap our baseline model called RefCLIP [10], which formulates one-stage weakly supervised REC as an anchor-text matching problem.

In particular, given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a textual expression $T \in \mathbb{R}^{L}$, RefCLIP firstly extracts anchor features $f_a \in \mathbb{R}^{n\times d}$ and text features $f_t \in \mathbb{R}^{l\times d}$ from the pre-trained YOLOv3 [40] and an LSTM [7], respectively. Notably, these anchor features can be decoded to bounding boxes $B \in \mathbb{R}^{n\times 4}$ via the detection head of YOLOv3. This process can be formulated by

$$a^* = \arg\max_{f_a \in F_v} \phi(f_a, f_t), \tag{1}$$

where $F_v \in \mathbb{R}^{(h\times w)\times d}$ denotes the visual features. $\phi(\cdot)$ refers to the matching function. After that, RefCLIP can obtain the bounding box of the referent by

$$b^* = \mathcal{D}(a^*). \tag{2}$$

$\mathcal{D}(\cdot)$ denotes the detection head and $b^* \in \mathbb{R}^4$ is the predicted bounding box, respectively. To achieve weakly supervised learning, anchor-text contrastive objective is adopted for optimization, which can be written by

$$\mathcal{L}_c = -\log \frac{\phi(f_a^i, f_t^i)/\tau}{\sum_{j=1}^{N} \phi(f_a^j, f_t^i)/\tau}. \tag{3}$$

Here, N and τ denote the batch size and the contrastive temperature, respectively.

As defined in Eq. 1 and 3, the performance upper bound of RefCLIP is highly dependent on the anchor-text matching, *i.e.*, $\phi(f_a, f_t)$. However, recent anchor features often fall short in describing fine-grained visual information, thus hindering accurate vision-language alignment.

4. DViN

To address above limitations, we propose a novel framework for weakly supervised REC, namely *Dynamic Visual routing Network* (DViN). As shown in Fig. 3, DViN overcomes the visual shortcoming of weakly supervised REC from the perspective of feature combination and alignment. In DViN, we incorporate various advanced visual encoders

into RefCLIP through a sparse routing mechanism, *e.g.,* CLIP [39] and DINO [2]. This additional visual knowledge greatly complements the limited visual capability of the default RefCLIP. To approach this target, DViN includes a novel Dynamic Visual Routing (DVR) strategy to dynamically combine visual features in an efficient way. Meanwhile, a novel Routing-based Feature Alignment (RFA) is used to further facilitate the vision-language alignment.

4.1. Dynamic Visual Routing

As shown in Fig. 3, DVR formulates the mixture of visual features as a sparse routing process, which only activates two visual experts during one forward process. In this case, DVR can not only best meet the visual requirement of weakly supervised REC, while still maintaining the efficiency. Specifically, DVR can be formulated by

$$a^* = \underset{f_{a0}\in F_{v0},\ldots,f_{am}\in F_{vm}}{\arg\max} \phi(\sum_{j\in S_e} f_{aj} r_j, f_t), \tag{4}$$

where r_j denote the routing weights generated by the router $\mathcal{R}(\cdot)$. S_e is the index of the selected visual features. Notably, the number of S_e is set to 2 to maintain the routing sparsity. In Eq. 4, the key to DVR lies in the router design and the routing loss, which determines whether the key visual features can be accurately selected and combined.

Vision-conditioned router. Vision-conditioned router (VCR) aims to dynamically produce routing weights to select the top-*k* visual experts from a pool of available models, *e.g.,* CLIP [39], DINOv2 [37], and SAM [14]. In VCR, routing weights are dependent on the given visual features, *i.e.,* features of DarkNet, which can be formulated as:

$$\mathcal{R}(F_{v0}) = \text{softmax}(\sum_{i=0}^{h\times w} F_{v0}^i W_r + b_r), \tag{5}$$

where $W_r \in \mathbb{R}^{d\times k}$ and $b_r \in \mathbb{R}^k$ denote the weight and bias, respectively. Then, the routing process can be defined by

$$S_e = \{0\} \cup \{\arg\max_j(\mathcal{R}(F_{v0})_j)\} \tag{6}$$

In Eq. 6, F_{v0} refers to visual features of DarkNet and we always keep it selected. To explain, at the beginning of training, the router tends to randomly select visual features to combine, thus harming the stability of training. To alleviate this issue, we keep visual features of DarkNet always selected, thus stabilizing the feature space during the initial training. Therefore, the routing is equivalent to making a top-1 selection from the remaining visual encoders.

Diversity-aware routing learning. To maximize the benefits of VCR during training, we introduce a diversity-aware loss to encourage the selection probabilities more evenly among the available experts, which is defined by

$$\mathcal{L}_{\text{r}} = \sum_{i=1}^{m} r_i \log(r_i). \tag{7}$$

Here, r_i is the routing score. With this loss, we can balance the load among the experts and prevent any single expert from being over-utilized.

4.2. Routing-based Feature Alignment

Based on DVR, we introduce Routing-based Feature Alignment (RFA) to further improve the routed visual representation during the weakly supervised optimization.

Text-guided alignment. The target of text-guided alignment aims to enrich visual semantics via the learning from textual information. To achieve this, we conduct the feature alignment between the routed anchor features and the text features. In particular, we first adopt an MLP layer to project anchor features to the same dimension as the text features, which can be written by

$$\hat{f}_a = \text{MLP}(\sum_{j\in S_e} f_{aj} r_j). \tag{8}$$

Then, the alignment loss is computed between anchor and text features by

$$\mathcal{L}_{\text{dt}} = \left(\hat{f}_a - f_t\right)^2. \tag{9}$$

Vision-guided alignment. Different from text-guided alignment, vision-guided alignment aims to distill the visual knowledge from the combined anchor feature to the individual anchor feature. Specifically, we first project the combined feature and the individual feature into a shared embedding space by

$$\begin{aligned} f'_a &= \text{MLP}(\sum_{j\in S_e} f_{aj} r_j), \\ f'_{a0} &= \text{MLP}(f_{a0}). \end{aligned} \tag{10}$$

In this process, we hope the individual feature that can represent the visual semantics of the combined feature, while also being distinct from others. Therefore, vision-guided alignment loss is defined as a contrastive function by

$$\mathcal{L}_{\text{dv}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}\left(f_a'^{(i)}, f_{a0}'^{(i)}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(f_a'^{(i)}, f_{a0}'^{(j)}\right)/\tau\right)}. \tag{11}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, N is the batch size, and τ is a temperature hyperparameter.

4.3. Overall Loss

Based on Eq. 3, 10 and 12, the final weakly supervised loss can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{c}} + \lambda_{\text{r}}\mathcal{L}_{\text{r}} + \lambda_{\text{dt}}\mathcal{L}_{\text{dt}} + \lambda_{\text{dv}}\mathcal{L}_{\text{dv}}, \tag{12}$$

where λ_{r}, λ_{dt} and λ_{dv} are hyperparameters that control the contributions of distillation loss term.

Table 1. **Comparison of different mixtures of visual features (MoVs) and DViN on RefCOCO and RefCOCO+.** All experiments are conducted on RefCLIP. DN, DViT, SViT, CViT, CCNN refer to DarkNet, DINOv2-ViT, SAM-ViT, CLIP-ViT and CLIP-CNN for short, respectively. Inference speed is tested on an NVIDIA 3090 GPU.

MoV strategy	Dynamic Routing	RefCOCO			RefCOCO+			Inference Speed
		val	test A	test B	val	test A	test B	
DarkNet (Baseline)	×	60.34	58.75	57.01	40.36	39.99	39.33	31.3fps
DINOv2-ViT	×	55.19	59.80	43.65	45.92	52.02	34.79	27.8fps
SAM-ViT	×	36.40	38.44	31.57	31.94	33.09	29.84	33.4fps
CLIP-ViT	×	40.99	46.08	32.99	40.36	45.33	32.29	28.5fps
CLIP-CNN	×	40.68	46.82	32.48	42.51	48.60	33.44	20.0fps
DN+DViT	×	63.23	66.59	52.32	49.12	54.87	38.96	26.5fps
DN+DViT+SViT	×	61.95	65.79	51.89	48.76	54.43	38.22	25.6fps
DN+DViT+SViT+CViT	×	61.67	65.91	52.07	48.63	55.64	38.90	20.1fps
DN+DViT+SViT+CViT+CCNN	×	62.14	65.26	52.89	48.95	54.06	38.90	14.5fps
DViN (ours)	✓	67.67	70.90	59.39	52.54	57.52	45.31	29.8fps

Table 2. **Cumulative ablation study of DViN on RefCOCO and RefCOCO+.** "V-Router" refers to our dynamic visual router. "R-loss", "T-loss" and "V-loss" denote the routing loss, text-guided alignment loss and vision-guided alignment loss, respectively.

Method	RefCOCO		RefCOCO+		Inference Speed
	test A	test B	test A	test B	
RefCLIP	58.75	57.01	39.99	39.33	31.3 fps
+ V-Router	68.68	57.03	56.84	43.50	29.8 fps
+ R-loss	69.11	59.01	56.96	43.85	29.8 fps
+ T-loss	70.30	59.01	57.19	44.24	29.8 fps
+ V-loss	70.90	59.39	57.52	45.31	29.8 fps

Table 3. **Comparison of different routing conditions on RefCOCO and RefCOCO+.**

Method	RefCOCO		RefCOCO+	
	test A	test B	test A	test B
Text	67.13	56.17	54.90	43.07
Image	70.90	59.39	57.52	45.31
Image+text	67.40	58.50	56.04	43.13

Table 4. **Ablations of visual encoders' number in DViN.**

Top-N	RefCOCO		RefCOCO+		Inference Speed
	test A	test B	test A	test B	
Top-1	70.90	59.39	57.52	45.31	29.8fps
Top-2	70.03	58.48	56.98	42.62	24.5fps
Top-3	68.44	57.25	56.75	42.68	16.7fps
Top-4	67.33	56.03	56.18	40.88	15.8fps

5. Experiments

5.1. Datasets and Metrics

RefCOCO [35] has 142,210 referring expressions and 50,000 objects in 19,994 images from the MSCOCO dataset [35]. The expressions in RefCOCO mostly describe the absolute position of objects. RefCOCO+ [35] includes 141,564 referring expressions for 49,856 bounding boxes in 19,992 MSCOCO images. Unlike RefCOCO, RefCOCO+ contains more descriptions of relationships and attributes. RefCOCOg [34] has 104,560 referring expressions and 54,822 bounding boxes for 26,711 images, with longer and more complex expressions compared to RefCOCO and RefCOCO+. ReferItGame [12] has 120,072 referring expressions for 99,220 bounding boxes in 19,997 images.

For REC task, we use IoU@0.5 as our metric. This means a prediction is deemed accurate if the intersection over union (IoU) with the ground truth exceeds 0.5. For RES task, we adopt the mIoU metric for evaluation, which averages the IoU scores across all test samples.

5.2. Implementation Details

As in RefCLIP, the image resolution is fixed at 416×416. The maximum text length is set to 15 for RefCOCO, 15 for RefCOCO+, and 20 for RefCOCOg. By default, we use YOLOv3 [40] as the detection network. For weakly supervised RES, the detection network is YOLOv5-Seg [11]. The additional visual encoders in DViN include CLIP-ViT [39], DINOv2-ViT [37], EfficientSAM-ViT [50], and CLIP-Convnext [43]. During weakly supervised training, we use the Adam optimizer with a learning rate of 1e-4 and a batch size of 64. The λ of diversity-aware loss is set to 0.01. The training process lasts for 25 epochs with a cosine decay schedule. Other settings follow those of RefCLIP and experiments are conducted on NVIDIA 3090 GPUs.

Table 5. **Comparison with state-of-the-art methods on four REC benchmark datasets.** "*GT proposals*" refers to using the official MSCOCO annotations as candidate regions. "*Pseudo Label*" indicates that the student REC model is trained with 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 [36]$_{\text{CVPR'18}}$	-	33.29	30.13	-	34.60	31.58	30.26	-	-
ARN [24]$_{\text{ICCV'19}}$	38.05	36.43	36.47	34.53	36.40	36.12	39.62	-	-
KPRN [25]$_{\text{MM'19}}$	36.34	35.28	37.72	37.16	36.06	39.29	38.37	33.87	-
DTWREG [47]$_{\text{TPAMI'21}}$	39.21	41.14	37.72	39.18	40.01	38.08	43.24	-	-
EARN [27]$_{\text{TPAMI'22}}$	38.08	38.25	38.59	37.54	37.58	37.92	45.33	36.86	-
Det Proposals:									
VC[36]$_{\text{CVPR'18}}$	-	32.68	27.22	-	34.68	28.10	29.65	14.50	-
KAC Net [3]$_{\text{CVPR'18}}$	-	-	-	-	-	-	-	15.83	-
MATN [54]$_{\text{CVPR'18}}$	-	-	-	-	-	-	-	13.61	-
ARN [24]$_{\text{ICCV'19}}$	32.17	35.25	30.28	32.78	34.35	32.13	33.09	26.19	5.7fps
IGN [53]$_{\text{NeurIPS'20}}$	34.78	37.64	32.59	34.29	36.91	33.56	34.92	-	-
DTWREG [47]$_{\text{TAPAMI'21}}$	38.35	39.51	37.01	38.91	39.91	37.09	42.54	-	5.9fps
RelR [28]$_{\text{CVPR'21}}$	-	-	-	-	-	-	-	37.68	-
NCE+Dist [49]$_{\text{CVPR'21}}$	-	-	-	-	-	-	-	38.39	-
RefCLIP [10]$_{\text{CVPR'23}}$	60.36	58.58	57.13	40.39	40.45	38.86	47.87	39.58	31.3fps
APL [32]$_{\text{ECCV'24}}$	64.51	61.91	63.57	42.70	42.84	39.80	50.22	**41.80**	26.7fps
DViN (ours)	**67.67**	**70.90**	**59.39**	**52.54**	**57.52**	**45.31**	**55.04**	40.63	29.8fps
Pesudo Labels:									
RefCLIP$_{\text{-SimREC}}$ [10]	62.57	62.70	61.22	39.13	40.81	36.59	45.68	42.33	54.8fps
RefCLIP$_{\text{-Transvg}}$ [10]	64.08	63.67	63.93	39.32	39.54	36.29	45.70	42.64	19.3fps
DViN$_{\text{-SimREC}}$ (ours)	**67.29**	**73.09**	60.65	**51.54**	**59.06**	**39.59**	**51.73**	**45.43**	38.4fps
DViN$_{\text{-Transvg}}$ (ours)	64.99	68.87	**64.48**	50.72	57.36	38.64	50.47	44.26	19.3fps

5.3. Experimental Results

Ablation studies. We conduct extensive experiments to validate the designs of DViN in Tab. 1, 2, 3 and 4. In particular, Tab. 1 compares different MoV baselines with DViN. For this table, the first observation is that most pre-trained visual features are still inferior to the default DarkNet [40] on REC tasks, *e.g.,* -5.15% of DINOv2-ViT [37] on RefCOCO *val*. However, when mixing DINOv2-ViT with the DarkNet, the performance boosts from 55.19% to 66.59%, suggesting the effectiveness of simple MoVs. Besides, we also observe that the simple MoVs easily reach their performance upper bound. For example, "DN+DViT" achieves similar performance to "DN+DViT+SViT+CViT+CCNN" on RefCOCO. Meanwhile, directly increasing the number of visual features does not always improve performance, *e.g.,* -0.32% after mixing SViT. More importantly, the simple MoV baseline will obviously reduce the inference speed due to the additional visual encoders, *e.g.,* up to -16.8 fps. These results confirm the effectiveness of simple MoVs while also indicating their shortcomings in performance and efficiency. In contrast, with the novel visual routing design, DViN not only achieves the best performance against all baselines, but also maintains the promising inference efficiency. These results extensively validate our motivation and designs of DViN. Notably, the inclusion of VTR enhances DViN's cross-modal correspondence, boosting its performance on RefCOCO+ *testA* by 1.19%, with an additional 0.60% improvement from CFA on top of VTR.

In Table 2, we present a cumulative ablation study to assess the contributions of each design element in DViN. Specifically, vision-conditioned routing, which dynamically integrates visual features into RefCLIP, achieves a substantial performance boost of up to +9.93% on RefCOCO+. Additionally, our Diversity-aware routing learning (R-loss), which promotes balanced utilization of visual features across experts, provides a further +0.43% gain. Moreover, the Text-guided alignment loss (T-loss) enhances alignment with textual semantics, resulting in an additional +1.19% improvement on RefCOCO test A. The Vision-guided alignment loss (V-loss) further refines visual representation by distilling visual knowledge into individual

Table 6. **Comparison of DViN and existing methods on weakly supervised RES.**

Method	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val-g
AMR [38]$_{\text{AAAI'22}}$	14.12	11.69	17.47	14.13	11.47	18.13	15.83
GroupViT [51]$_{\text{CVPR'22}}$	18.03	18.13	19.33	18.15	17.65	19.53	19.97
CLIP-ES [21]$_{\text{CVPR'23}}$	13.79	15.23	12.87	14.57	16.01	13.53	14.16
GbS [1]$_{\text{ICCV'21}}$	14.59	14.60	14.97	14.49	14.49	15.77	14.21
WWbL [44]$_{\text{NeurIPS'22}}$	18.26	17.37	19.90	19.85	18.70	21.64	21.84
TSEG [45]$_{\text{arXiv'20}}$	30.12	-	-	25.95	-	-	22.62
ALBEF [17]$_{\text{NeurIPS'21}}$	23.11	22.79	23.42	22.44	22.07	22.51	24.18
I-Chunk [15]$_{\text{ICCV'23}}$	31.06	32.30	30.11	31.28	32.11	30.13	32.88
TRIS [23]$_{\text{ICCV'23}}$	31.17	32.43	29.56	30.90	30.42	30.80	36.00
PKS [16]$_{\text{arXiv'22}}$	49.27	52.23	45.64	37.79	42.09	32.87	36.43
APL [32]$_{\text{ECCV'24}}$	55.92	54.84	55.64	34.92	34.87	35.61	40.13
DViN (ours)	**61.43**	**63.81**	**56.97**	**46.79**	**51.87**	**39.85**	**46.49**

features, adding another +0.60% gain. Collectively, these components highlight the effectiveness of DViN's design in adapting to diverse visual and textual demands.

In Table. 3, we compare different routing conditions for the router to generate routing weights. The observation is that the router can produce better dynamic weights based on image-only conditions. When we use text and image+text as the routing conditions, performance drops by -1.98% and -1.98% on RefCOCO *text A*. These results suggest that the routing mainly relies on information from image features instead of text features. To explain, under the weakly supervised setting, the text features are not well aligned with visual features at the beginning of the training, which can not accurately determine the selection of visual features.

In Table 4, we compare the impact of varying the number of visual features in our model. Our first observation is that increasing the number of visual features does not always lead to better performance. Specifically, fewer features (top-1 and top-2) generally outperform those with more features (top-3 and top-4). For instance, in the RefCOCO dataset, the top-1 and top-2 settings achieve the highest accuracy on the *test B* and *test A* splits, respectively. Meanwhile, the inference speed typically declines with the increase in the number of features, *e.g.,* 29.8fps of top-1 *vs.* 15.8fps of top-4. This suggests that adding more features will often increase the computational overhead without proportionate gains in performance, which further confirms the sparse design of DViN.

Comparison with weakly supervised REC methods. In Table. 5, we compare DViN with a set of methods on four REC benchmark datasets. From this table, our approach significantly outperforms existing methods across multiple datasets and metrics. Notably, DViN achieves the highest accuracy in nearly all splits when compared with methods of both ground truth (GT) proposals and detection (Det) proposals *e.g.,* +6.01% on RefCOCO *val* set and +10.41% on *test A* against RefCLIP. Besides, we also observe that our methods show more gains on RefCOCO+, which has higher requirements for visual details. In addition to accuracy, the inference efficiency of DViN is also notable, which demonstrates a competitive speed of 29.8 fps against existing methods, *e.g.,* +23.9fps than DTWREG. These results also confirm the benefits of DViN to visual description ability. We further evaluate DViN under the pseudo-labeling setting, where DViN performs as the teacher to supervise common REC models. From these results, we can see that DViN can produce accurate pseudo-labels against RefCLIP, thus yielding better student performance.

Generalizations to weakly supervised RES. To validate the generalization ability of DViN, we apply it to the weakly supervised RES task. In particular, we use the YOLOv5-Seg [11] as the detection network. Therefore, we can directly extend the anchor-text matching framework of DViN to weakly supervised RES. As shown in Table. 6, DViN can also achieve the state-of-the-art performance on three benchmark datasets, *e.g.,* up to +31.38% gains on RefCOCO. Even compared to PKS [16], which uses additional annotations, DViN still demonstrates comprehensive advantages on all benchmark datasets. These obvious performance gains show the significance of visual description ability in weakly supervised RES and confirm the generalization ability of DViN.

Qualitative analysis. To gain deep insights into DViN, we visualize its predictions in Fig. 4. From Fig. 4 (a), we observe that visual features play unique roles in different samples. For example, SAM-ViT [14] is often selected when visual objects are large and salient. CLIP [39] and DINOv2 [37] excel at processing visual objects with rich colors and detailed appearance. These examples potentially indicate the functionality of DViN. Additionally, we compare the predictions of DViN and RefCLIP in Fig. 4 (b). From this figure, we can see that DViN shows significantly better

Figure 4. **Visualizations of DViN and RefCLIP on RefCOCO and RefCOCO+.** Sub-figure (a) shows that different samples have their preferences on visual encoders. Sub-figure (b) demonstrates the better visual ability of DViN than RefCLIP.

results on fine-grained expressions. For example, DViN can distinguish the subtle difference between two "vehicles" in Exp-6, while RefCLIP fails to locate the target one. Overall, these examples further illustrate the benefits of DViN. To gain further insights into DViN's limitations, we examine several failure cases in Fig. 4 (c). In Exp-3, DViN fails to locate the "bear behind the tree," incorrectly focusing on a foreground object. Similarly, in Exp-4, DViN struggles to distinguish the "larger piece of upside-down broccoli near the rice." These examples illustrate DViN's limitations in handling complex, fine-grained expressions, highlighting areas for improvement in precise target localization.

6. Conclusion

In this paper, we focus on weakly supervised referring expression comprehension (REC) and identify weak visual ability as its main performance limitation. We observe that a mixture of visual features significantly improves fine-grained object detection. Based on this, we propose a novel framework, Dynamic Visual Routing Network (DViN), which formulates visual feature extraction in weakly supervised REC as a dynamic routing process. Additionally, an innovative routing-based feature alignment enhances routing features via intra-modal and inter-modal alignment. Experimental results validate DViN's state-of-the-art performance and inference efficiency.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 623B2088) and the China Postdoctoral Science Foundation (No. 2024M761548).

References

[1] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1801–1812, 2021. 7

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[3] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4042–4050, 2018. 6

[4] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 1, 2

[5] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 3

[6] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023. 2

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[8] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 1, 2

[9] Inwoo Hwang, Hyeonwoo Kim, and Young Min Kim. Text2scene: Text-driven indoor scene stylization with part-aware details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1890–1899, 2023. 1

[10] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2681–2690, 2023. 1, 3, 6

[11] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022. 5, 7

[12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 5

[13] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 2

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4, 7

[15] Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21870–21881, 2023. 7

[16] Hui Li, Mingjie Sun, Jimin Xiao, Eng Gee Lim, and Yao Zhao. Fully and weakly supervised referring expression segmentation with end-to-end learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5999–6012, 2023. 7

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 7

[18] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1

[19] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[21] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 7

[22] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 1, 2

[23] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023. 7

[24] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019. 1, 6

[25] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019. 1, 6

[26] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959, 2019. 1, 2, 3

[27] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3003–3018, 2022. 6

[28] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5612–5621, 2021. 1, 3, 6

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[30] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020. 2

[31] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 2

[32] Yaxin Luo, Jiayi Ji, Xiaofu Chen, Yuxin Zhang, Tianhe Ren, and Gen Luo. Apl: Anchor-based prompt learning for one-stage weakly supervised referring expression comprehension. In *European Conference on Computer Vision*, pages 198–215. Springer, 2025. 6, 7

[33] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 259–274. Springer, 2020. 1

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 5

[35] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 2, 5

[36] Yulei Niu, Hanwang Zhang, Zhiwu Lu, and Shih-Fu Chang. Variational context: Exploiting visual and textual context for grounding referring expressions. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):347–359, 2019. 6

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 5, 6, 7

[38] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2117–2125, 2022. 7

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4, 5, 7

[40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 3, 5, 6

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[42] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 817–834. Springer, 2016. 2

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[44] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems*, 35:28222–28237, 2022. 7

[45] Robin Strudel, Ivan Laptev, and Cordelia Schmid. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*, 2022. 7

[46] Xiuchao Sui, Shaohua Li, Hong Yang, Hongyuan Zhu, and Yan Wu. Language models can do zero-shot visual referring expression comprehension. 2023. 2

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

[48] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[49] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021. 1, 3, 6

[50] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024. 5

[51] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 7

[52] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 1, 2

[53] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134, 2020. 3, 6

[54] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018. 1, 3, 6

[55] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):134–143, 2021. 2

[56] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 2