DARA: DOMAIN- AND RELATION-AWARE ADAPTERS MAKE PARAMETER-EFFICIENT TUNING FOR VISUAL GROUNDING

Ting Liu[1]* Xuyang Liu[2]* Siteng Huang[3] Honggang Chen[2] Quanjun Yin[1]
Long Qin[1] Donglin Wang[3] Yue Hu[1†]

[1]College of Systems Engineering, National University of Defense Technology, Changsha, China
[2]College of Electronics and Information Engineering, Sichuan University, Chengdu, China
[3]School of Engineering, Westlake University, Hangzhou, China
liuting20@nudt.edu.cn, liuxuyang@stu.scu.edu.cn

ABSTRACT

Visual grounding (VG) is a challenging task to localize an object in an image based on a textual description. Recent surge in the scale of VG models has substantially improved performance, but also introduced a significant burden on computational costs during fine-tuning. In this paper, we explore applying parameter-efficient transfer learning (PETL) to efficiently transfer the pre-trained vision-language knowledge to VG. Specifically, we propose **DARA**, a novel PETL method comprising **D**omain-aware **A**dapters (DA Adapters) and **R**elation-aware **A**dapters (RA Adapters) for VG. DA Adapters first transfer intra-modality representations to be more fine-grained for the VG domain. Then RA Adapters share weights to bridge the relation between two modalities, improving spatial reasoning. Empirical results on widely-used benchmarks demonstrate that DARA achieves the best accuracy while saving numerous updated parameters compared to the full fine-tuning and other PETL methods. Notably, with only **2.13%** tunable backbone parameters, DARA improves average accuracy by **0.81%** across the three benchmarks compared to the baseline model. Our code is available at https://github.com/liuting20/DARA.

Index Terms— Visual grounding, parameter-efficient transfer learning, vision-language models

1. INTRODUCTION

Visual grounding (VG) [1, 2] is one of the most challenging tasks in the vision-language learning fields. Compared with tasks like text-image retrieval and image captioning, VG requires more fine-grained vision-language alignment for spatial reasoning to accurately locate an object described by a language expression. Existing VG methods [3, 4, 5, 6] follow a standard *pretrain-then-finetune* paradigm, which first pre-trains the vision and language backbones on large-scale datasets and then fine-tunes the whole models on downstream
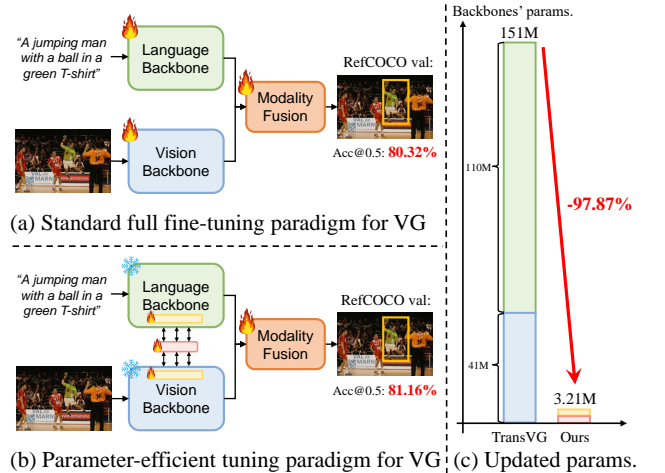
---

*Equal contribution. †Corresponding author.

Fig. 1. Comparison of (a) full fine-tuning [5] and (b) our PETL method for visual grounding. (c) Freezing (❄) the pre-trained backbones and updating (🔥) our DARA reduces **97.87%** of backbone updated parameters while achieving even **stronger** performance than full fine-tuning paradigm.

specific datasets [1, 2]. As the backbone networks of VG models become increasingly deeper, the number of parameters grows rapidly, from 47M in MAttNet [3] to 151M in TransVG [5]. Fine-tuning such large VG models can be computationally expensive and time-consuming in practice.

Recently, a variety of parameter-efficient transfer learning (PETL) methods [7, 8, 9] have demonstrated high efficiency and extensibility in transferring diverse knowledge from pretraining to downstream tasks. Specifically, these PETL methods freeze nearly all parameters of the pre-trained model, and only update a small number of additional adaptable parameters. With just a tiny fraction of trainable parameters, they can achieve comparable performance to full fine-tuning. Current PETL methods are proposed to address single-modality tasks or simple cross-modal tasks, such as language understanding [7], image classification [8] and text-image/video retrieval [10, 11]. This gives rise to an intriguing question: *can*

these PETL methods be directly applied to the visual grounding task to reduce the computational and time cost of fine-tuning?. Experiments in Section 4.2 demonstrate that directly applying them to the VG task can lead to considerable performance degradation. We suppose that the PETL methods lack sufficient interaction between the visual and linguistic representations for spatial reasoning.

Motivated by the above analysis, this paper aims to effectively and efficiently fine-tunes a standard transformer-based VG model to match the full fine-tuning results but with remarkably decreased computational cost. Specifically, we propose a novel adapter-based PETL method termed **DARA** that incorporates **D**omain-aware **A**dapters (DA Adapters) and **R**elation-aware **A**dapters (RA Adapters), for visual grounding. DA Adapters first transfer the pre-trained intra-modality representations in the dual backbones to more fine-grained vision and language representations to the VG **domain**. Subsequently, RA Adapters enable early cross-modal interactions by sharing the adapters' weights between the two backbones, thus bridging the **relation** between the two modalities which facilitates the cross-model reasoning abilities of the baseline model. DARA based on a pre-trained VG method, TransVG [5], can reduce approximately **97.87%** updated parameters of the pre-trained vision and language backbones while still achieving comparable performance with the full fine-tuning paradigm, as shown in Figure 1. Furthermore, we surprisingly discover that the early cross-modal interactions enabled by our proposed RA Adapters can improve fusion effectiveness for spatial reasoning, which is validated by the experiments depicted in Figure 3.

Our main contributions can be summarized as threefold:

- We present an in-depth exploration of adapting parameter-efficient transfer learning (PETL) methods for visual grounding. To the best of our knowledge, this is the first study to investigate parameter-efficient tuning paradigm for visual grounding.
- We propose DARA, a novel PETL framework incorporating two tailored adapters: Domain-aware Adapters and Relation-aware Adapters, to facilitate effective and efficient intra- and inter-modality representation transfer for the visual grounding task.
- Extensive experiments on three widely-used VG benchmarks demonstrate that DARA can achieve the best accuracy while saving numerous updated parameters compared to full fine-tuning and other PETL methods.

2. RELATED WORK

Visual Grounding. Visual grounding (VG) [1, 2] is a challenging vision-language modeling task that aims to localize the specific object in an image by the given textual description. Currently, most VG methods can be grouped into three types based on their network architectures: (1) **Two-stage** VG methods [3, 12, 13] first employ a pre-trained object de-

tector to generate series of sparse region proposals, and then select the most similar one to the description for grounding. (2) **One-stage anchor-based** VG methods [4, 14, 15] integrate detect-then-select progress into one single pipeline that jointly fuses visual and textual representations at the intermediate layers of the detector and predicts the box by the pre-defined anchors. (3) **Transformer-based** VG methods [5, 16, 6] are proposed for end-to-end VG, benefiting from their strong intra- and inter-modality representation learning. Unfortunately, fine-tuning recent VG methods is becoming far more computationally expensive and time-consuming.

Parameter-efficient Transfer Learning. Parameter-efficient transfer learning (PETL) [7, 9] has become a prevalent approach for efficiently transferring fixed pre-trained models to downstream tasks by fine-tuning just a few parameters. PETL methods can be broadly categorized into two types: (1) Updating additional parameters in the module inserted into the model [7, 8] or appended with the input data [17, 18]; (2) Decomposing the weight matrices of pre-trained models into two low-rank matrices and only updating the small factorization matrices [9]. Recently, adapter-based PETL methods have shown strong efficiency, extensibility, and performance in various downstream vision-language tasks [11, 10]. Our work focus on efficiently fine-tuning a pre-trained transformer-based VG method by carefully designed two types of adapters, which facilitate efficient intra- and inter-modality representation transferring.

3. METHODOLOGY

3.1. Baseline Model

As depicted in Figure 2 (a), we follow the architecture and training objective of TransVG [5], a typical transformer-based VG method, to implement our model. It consists of three components: (1) Vision Backbone, (2) Language Backbone, and (3) Vision-language Transformer.

Vision Backbone. The Vision Backbone, comprising a pre-trained ResNet [19] and a vision transformer, encodes the image into a high-quality vision embedding. Specifically, given an input image $z_0 \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the Vision Backbone first uses the ResNet to generate a 2D feature map $z \in \mathbb{R}^{H \times W \times C}$, where $C = 2048$ is the channel dimension, $H = \frac{H_0}{32}$, and $W = \frac{W_0}{32}$. A 1×1 convolutional layer then reduces C to $C_v = 256$, producing $z' \in \mathbb{R}^{H \times W \times C_v}$. After flattening z' into the vision embedding $z_v \in \mathbb{R}^{N_v \times C_v}$, where $N_v = H \times W$ represents the number of input tokens, z_v added with the position embedding is input to the vision transformer including 6 stacked transformer encoder layers. The vision transformer performs parallel processing on z_v and outputs an enhanced vision embedding $f_v \in \mathbb{R}^{N_v \times C_v}$.

Language Backbone. The Language Backbone, adopting a pre-trained BERT [20], encodes the language expression into a language embedding. Specifically, given a language expres-
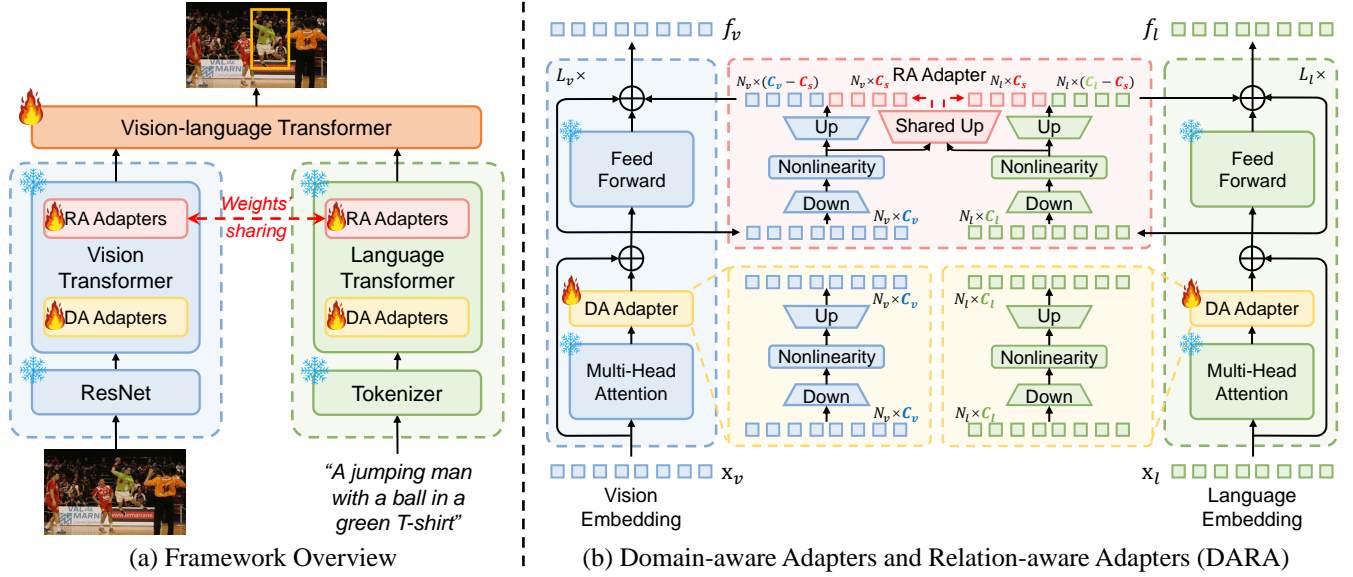
Fig. 2. (a) Overview of our proposed parameter-efficient tuning framework for visual grounding. We freeze (❄) the vision and language backbones and update (🔥) our DARA, comprising the Domain-aware Adapters (DA Adapters) and Relation-aware Adapters (RA Adapters). (b) Detailed design of the DARA. The DA Adapters transfer the pre-trained rich intra-modality representations, making them more fine-grained for the visual grounding domain. Subsequently, the RA Adapters then share the adapters' weights to bridge the relation between the two backbones and capture inter-modality representations.

sion, the Language Backbone first converts each word ID into a one-hot vector, then tokenizes these vectors into series of language tokens. The language tokens, concatenated with a `[CLS]` token at the beginning and a `[SEP]` token at the end, are input to the language transformer comprising 12 identical transformer encoder layers. Similar to the vision transformer, the language transformer outputs an enhanced language embedding $\boldsymbol{f}_l \in \mathbb{R}^{N_l \times C_l}$, where N_l and C_l represent the number and channel dimension of language tokens, respectively.

Vision-language Transformer. The Vision-Language Transformer (V-L Transformer) aims to fuse the two modality embeddings and output a `[REG]` token for generating the bounding box. Specifically, after the vision and language backbones output the enhanced vision embedding $\boldsymbol{f}_v \in \mathbb{R}^{N_v \times C_v}$ and language embedding $\boldsymbol{f}_l \in \mathbb{R}^{N_l \times C_l}$, these embeddings are first projected into the same channel dimension $C_p = 256$. The joint embeddings $\boldsymbol{f}'_v \in \mathbb{R}^{N_v \times C_p}$ and $\boldsymbol{f}'_l \in \mathbb{R}^{N_l \times C_p}$ are then concatenated with a learnable `[REG]` token, forming the input $\boldsymbol{x}_0 \in \mathbb{R}^{(N_v + N_l + 1) \times C_p}$ for the V-L Transformer with a stack of 6 transformer encoder layers. Sequentially, the V-L Transformer outputs the `[REG]` token to a Prediction Head, implemented as a MLP with two 256-dimensional hidden layers with ReLU activation and a linear output layer, to regress the 4-dimensional box coordinates for the referred object.

3.2. Domain-aware and Relation-aware Adapters

As the vision and language backbones contain most model parameters (around 95%) and have acquired rich knowledge during pre-training, we attempt to freeze them during fine-

tuning. However, fully freezing the backbones deteriorates visual grounding performance, as shown in Table 2. We hypothesize that is because the pre-trained backbones lack **two key abilities** required for adapting to the visual grounding task: (1) Fine-grained intra-modality representations for precisely describing and locating the objects. (2) Early cross-modal interactions between the vision and language backbones to enable joint spatial reasoning. Based on the above analysis, we propose DARA with Domain-aware Adapters and Relation-aware Adapters for visual grounding, as shown in Figure 2. Note that we do not show the LayerNorm for simplicity.

DA Adapters. The core function of DA Adapters is to transfer the pre-trained rich and general intra-modality representations to more fine-grained representations for the VG domain. Specifically, the architecture of the DA Adapters follows the design of Adapters [7] in natural language processing (NLP), comprising a down-projection layer, a non-linear layer, and an up-projection layer. The DA Adapters are inserted after the Multi-Head Attention layers (MHA) in the vision and language transformers. Taking the DA Adapters in the vision transformer as an example, formally, given the visual feature map $\boldsymbol{x}_v \in \mathbb{R}^{N_v \times C_v}$ processed by the MHA, the function of DA Adapter in the transformer encoder can be expressed as:

$$\text{DA Adapter}(x_v) = x_v + s \cdot \sigma(x_v \mathbf{W}_{\text{down}}) \mathbf{W}_{\text{up}}, \quad (1)$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{C_v \times C_d}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{C_d \times C_v}$ are the weights of down- and up-projection layers, $\sigma(\cdot)$ is the ReLU activation function, and s is the scaling factor. In this way, DA Adapters can refine the rich pre-trained vision and lan-

guage representations into more fine-grained intra-modality representations for the VG domain during fine-tuning.

RA Adapters. Inspired by the existing efforts [11], our RA Adapters are designed to leverage the inter-modality representations, enabling early interaction between the two separate backbones, as shown in Figure 2 (b). Specifically, the RA Adapter is composed of a down-projection layer, a non-linear layer, a unique up-projection layer, and a shared up-projection layer. Unlike the CM Adapters [11], our RA Adapters are connected **in parallel** with the Feed Forward Networks (FFN) in both vision and language backbones to enhance the fine-tuning effectiveness, as evidenced in Table 4. Once the DA Adapters in both transformer encoders generate the enhanced embeddings, these embeddings are added with the original ones through skip-connections. Formally, given these adapted vision embedding $\boldsymbol{x}_v \in \mathbb{R}^{N_v \times C_v}$ and language embedding $\boldsymbol{x}_l \in \mathbb{R}^{N_l \times C_l}$, the down-projection layers first down-sample them to obtain the bottleneck features $\boldsymbol{z}_v \in \mathbb{R}^{N_v \times C_d}$ and $\boldsymbol{z}_l \in \mathbb{R}^{N_l \times C_d}$ as follows:

$$z_v = \sigma(x_v \mathbf{W}_{\text{down,vision}}), \tag{2}$$

$$z_l = \sigma(x_l \mathbf{W}_{\text{down,language}}), \tag{3}$$

where $\mathbf{W}_{\text{down,vision}} \in \mathbb{R}^{C_v \times C_d}$, $\mathbf{W}_{\text{down,language}} \in \mathbb{R}^{C_l \times C_d}$, and $\sigma(\cdot)$ is the ReLU activation function. Then, the up-projection layers with the sharing mechanism up-sample these bottleneck features to obtain the inter-modality features $\boldsymbol{f}_v \in \mathbb{R}^{N_v \times C_v}$ and $\boldsymbol{f}_l \in \mathbb{R}^{N_l \times C_l}$ as follows:

$$f_v = \text{Concat}[z_v \mathbf{W}_{\text{up,vision}}, z_v \mathbf{W}_{\text{up,share}}], \tag{4}$$

$$f_l = \text{Concat}[z_l \mathbf{W}_{\text{up,language}}, z_l \mathbf{W}_{\text{up,share}}], \tag{5}$$

where $\mathbf{W}_{\text{up,vision}} \in \mathbb{R}^{C_d \times (C_v - C_s)}$ and $\mathbf{W}_{\text{up,language}} \in \mathbb{R}^{C_d \times (C_l - C_s)}$ represent the weights of the unique vision and language up-projection layers, while $\mathbf{W}_{\text{up,share}} \in \mathbb{R}^{C_d \times C_s}$ and C_s are the weights and dimension of the shared up-projection layer. With this design, RA Adapters enable early cross-modal interactions between the two backbones.

As shown in Figure 2, freezing the pre-trained backbones and updating DARA can effectively fine-tune the baseline model with lower computation cost. Additionally, this pipeline facilitates fine-grained intra-modality representations and early inter-modality interactions.

4. EXPERIMENTS

4.1. Experimental Settings

Datasets and Evaluation Metrics. We conduct all experiments on three widely-used VG benchmarks: RefCOCO [1], RefCOCO+ [1] and RefCOCOg [2]. RefCOCO has 120,624 training, 10,834 validation, and 5,657 and 5,095 images for testA and testB. RefCOCO+ consists of 120,624 training, 10,758 validation, and 5,726 and 4,889 images for testA and

testB. RefCOCO+ includes some absolute-location expressions, making it more challenging. RefCOCOg has 25,779 images with 49,856 referred objects, which is divided into RefCOCOg-google (val-g) and RefCOCOg-umd (val-u and test-u). It provides longer and more complex expressions. We use Accuracy@0.5 to evaluate the performance.

Implementation Details. We adopt the backbone and encoder of the pre-trained DETR [21] as the vision backbone, and the pre-trained BERT-base [20] as the language backbone. The V-L Transformer is initialized with Xavier initialization. We initialize the adapters using Kaiming normal initialization, and insert them into the transformer encoder layers at the same indices as those in the DETR and BERT. The bottleneck dimensions of the DA Adapters are 128, the weight-sharing dimension C_s of the RA Adapters is 256, and the adapters' scaling factor s is 0.1. We fine-tune the network for 90 epochs, using the AdamW optimizer with a learning rate of 10^{-4} for V-L Transformer, 10^{-5} for the two backbones, and set the weight decay to 10^{-4}. The learning rate will be dropped by a factor of 10 after 60 epochs. For fair comparisons, all PETL methods follow the TransVG [5] as the baseline model, with the vision and language backbones frozen while the V-L Transformer is updated during fine-tuning.

4.2. Main Results

The main experimental results are presented in Table 1, from which we can observe that: (1) DARA achieves the best accuracy while ensuring parameter efficiency among all methods, thus validating its effectiveness and efficiency. (2) Compared with the intra-modality efficient transferring methods such as Adapter [7] and AdaptFormer [8], our method also achieves inter-modality efficient transferring, which brings significant performance improvements. (3) Compared with the most similar methods CM Adapter [11] and MRS-Adapter [10], our method considers both intra- and inter-modality knowledge transfer, thus achieving an average accuracy improvement of 4% on the three benchmarks. (4) DARA is the only PETL method that can achieve better performance than the baseline model [5], by updating only **2.13%** of the parameters in the vision and language backbones while achieving about **0.81%** accuracy improvement. This indicates that by combining intra- and inter-modality representation transferring, the PETL method has the potential to surpass the baseline model performance to some extent. Notably, DARA does better in complex expressions, since it outperforming TransVG [5] on all metrics in the more complex RefCOCO+ [1] and RefCOCOg [2], which contain more objects and more intricate language descriptions, compared to RefCOCO [1].

4.3. Ablation Study and Analysis

In this subsection, we analyze the effects of the two adapters and reasonable ways to combine and insert them into the baseline model. By default, DA Adapters are inserted sequentially

Methods	Backbone (vision/language)	#Updated Params.	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val-g	RefCOCOg val-u	RefCOCOg test-u
Full fine-tuning methods											
Two-stage:											
MAttNet [3]	RN101/LSTM	47M	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
RvG-Tree [12]	RN101/LSTM	47M	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
One-stage:											
FAOA [4]	DN53/LSTM	43M	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.26
SAFF [14]	DN53/BERT	152M	79.26	81.09	76.55	64.43	68.46	58.43	-	68.94	68.91
PFOS [15]	DN53/BERT	152M	77.37	80.43	72.87	63.74	68.54	55.84	61.46	67.08	66.35
Transformer-based:											
VGTR [16]	RN50/LSTM	52M	78.70	82.09	73.31	63.57	69.65	55.33	62.88	65.62	65.30
DMRNet [6]	DN53/BERT	152M	76.99	79.71	72.67	61.58	66.60	54.00	-	66.03	66.70
TransVG† [5]	RN50/BERT	151M	80.32	82.67	**78.12**	63.50	68.15	55.63	66.56	67.66	67.44
Parameter-efficient fine-tuning methods											
Adapter [7]	RN50/BERT	3.27M	78.02	79.89	75.23	61.35	66.34	54.21	63.18	65.26	66.65
LoRA [9]	RN50/BERT	2.37M	77.57	78.22	73.37	61.24	66.53	53.95	64.27	67.36	66.43
AdaptFormer [8]	RN50/BERT	2.38M	76.32	77.16	73.94	60.96	65.19	53.88	61.81	65.44	64.37
CM Adapter [11]	RN50/BERT	3.27M	77.37	78.81	74.07	61.34	66.10	53.31	63.93	65.75	64.72
MRS-Adapter [10]	RN50/BERT	1.58M	77.14	77.80	74.80	61.13	66.38	53.13	63.07	66.46	65.16
DARA (ours)	RN50/BERT	3.21M	**81.16**	**82.76**	76.72	**65.58**	**69.83**	**57.22**	**67.21**	**69.22**	**67.67**
$\Delta_{baseline}$	RN50/BERT	2.13%	+0.84	+0.09	-1.40	+2.08	+1.68	+1.59	+0.65	+1.56	+0.23

Table 1. Comparison on the RefCOCO [1], RefCOCO+ [1] and RefCOCOg [2]. #Updated Params.: the tunable parameters in the vision and language backbones. † represents the baseline model TransVG [5] for all parameter-efficient fine-tuning methods. $\Delta_{baseline}$ is the performance gap between our DARA and its baseline model TransVG [5]. **Bold** denotes the best results.

DA Adapters	RA Adapters	Updated Params.	RefCOCO val	RefCOCO testA	RefCOCO testB
		0	72.72	73.33	71.27
✓		1.63M	77.81	79.30	75.35
	✓	1.58M	77.32	77.37	75.64
✓	✓	3.21M	**81.16**	**82.76**	**76.72**

Table 2. Ablations of two adapters on RefCOCO [1] dataset.

Multi-head Attention	Feed-forword Network	Updated Params.	RefCOCO val	RefCOCO testA	RefCOCO testB
DA	DA	3.26M	79.20	80.83	76.57
RA	RA	3.16M	78.64	80.23	76.55
RA	DA	3.21M	78.30	79.56	75.16
DA	RA	3.21M	**81.16**	**82.76**	**76.72**

Table 3. Comparison of different adapter combination forms on the RefCOCO [1] dataset.

DA Adapters	RA Adapters	RefCOCO val	RefCOCO testA	RefCOCO testB
parallel	-	76.32	77.16	73.94
sequential	-	77.81	79.30	75.35
sequential	sequential	78.76	80.25	74.90
sequential	parallel	**80.26**	**82.49**	**76.65**
Ours ($C_s = 64$)		79.00	80.15	75.96
Ours ($C_s = 128$)		80.26	82.49	76.65
Ours ($C_s = 256$)		**81.16**	**82.76**	**76.72**

Table 4. Comparison of different adapter insertion forms and the performance of RA Adapters with different weight-sharing dimensions on the RefCOCO [1] dataset.

after the MHA and RA Adapters are inserted in parallel with the FFN in the Transformer Encoder.

Contribution of Each Component. As shown in Table 2, freezing the two backbones leads to much greater performance degradation, compared with the PETL baselines. We also report the results of using only DA Adapters or only RA Adapters. The results indicate that adopting either of them can lead to an accuracy improvement of around 5% compared to freezing the two backbones. By combining the DA and RA Adapters, DARA presents superior performance which demonstrates the effectiveness and efficiency of our method.

Comparison of Different Combination Ways. As shown in the first two rows of Table 3, inserting the DA Adapters after the MHA and FFN achieves better performance than inserting the RA Adapters. This indicates that it is more important to first refine the intra-modality representations. Surprisingly, first inserting the RA Adapters in parallel with the MHA followed by the DA Adapters inserted after the FFN results in the worst performance. This demonstrates that initially sharing the adapters' weights may impair the rich pre-trained intra-modality representations. Consequently, forcibly transferring these impaired representations to the VG domain leads to worth performance. In contrast, DARA first transfers the intra-modality representations to be more fine-grained for VG, and then shares the adapters' weights to bridge the two modalities, thereby achieving the best performance.

Comparison of Different Insertion Forms. Motivated by AdaptFormer [8], we investigate **parallel** (inserted in parallel) versus **sequential** (inserted sequentially) insertion formulations. As the upper part of Table 4 shows, sequential DA Adapters outperform parallel design by effectively transferring the pre-trained intra-modality representations from the MHA for VG. Meanwhile, the parallel design of RA
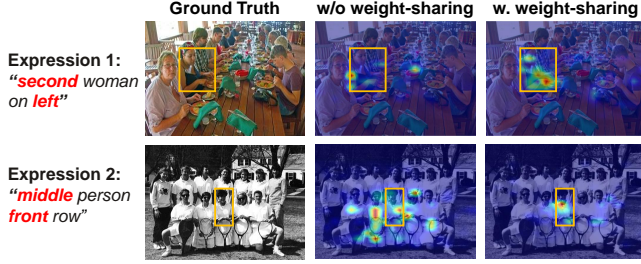
Fig. 3. Visualizations of cross-attention maps between the [REG] token and visual tokens under different strategies.

Adapters efficiently bridges the two backbones to capture the inter-modality representations, resulting in better performance. Furthermore, we explore the impacts of varying degrees of cross-modal interactions, shown in the lower part of Table 4. As CM Adapter [11] shows the best text-video retrieval performance by sharing half the visual dimensions, we find fully sharing the 256-dim visual features achieves the highest VG accuracy. We hypothesize that VG serves as a more challenging vision-language task, therefore, requiring much deeper cross-modal interactions.

Effectiveness of Weight-sharing Strategy. To further study the impact of early cross-modal interactions via RA Adapters, we compare DARA with a baseline without weight-sharing on challenging cases requiring spatial reasoning. We visualize the cross-attention maps between the [REG] token and visual tokens in the last layer of the V-L Transformer. As depicted in Figure 3, DARA's [REG] token focuses more on the referred object than the baseline. This suggests that weight-sharing strategy of RA Adapters promotes the synergy between vision and language modalities, thereby enhancing spatial reasoning for visual grounding.

5. CONCLUSION

In this paper, we present a new parameter-efficient tuning method DARA for visual grounding. DARA first utilizes Domain-aware Adapters to enhance the pre-trained intra-modality representations, making them fine-grained for the domain of VG. Sequentially, it employs Relation-aware Adapters to bridge the relation between the vision and language backbones and capture inter-modality representations. In this way, DARA achieves the best accuracy while saving numerous updated parameters across three benchmarks, and even surpasses the performance of the baseline model.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Licheng Yu, Patrick Poirson, Shan Yang, et al., "Modeling context in referring expressions," in *ECCV*, 2016.

[2] Junhua Mao, Jonathan Huang, Alexander Toshev, et al., "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016.

[3] Licheng Yu, Zhe Lin, Xiaohui Shen, et al., "MAttNet: Modular attention network for referring expression comprehension," in *CVPR*, 2018.

[4] Zhengyuan Yang, Boqing Gong, Liwei Wang, et al., "A fast and accurate one-stage approach to visual grounding," in *ICCV*, 2019.

[5] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, et al., "TransVG: End-to-end visual grounding with transformers," in *ICCV*, 2021.

[6] Zhipeng Zhang, Zhimin Wei, Zhongzhen Huang, et al., "One for all: One-stage referring expression comprehension with dynamic reasoning," *Neurocomputing*, 2023.

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al., "Parameter-efficient transfer learning for nlp," in *ICML*, 2019.

[8] Shoufa Chen, Chongjian Ge, Zhan Tong, et al., "AdaptFormer: Adapting vision transformers for scalable visual recognition," in *NeurIPS*, 2022.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, et al., "LoRA: Low-rank adaptation of large language models," in *ICLR*, 2022.

[10] Yuan Yuan, Yang Zhan, and Zhitong Xiong, "Parameter-efficient transfer learning for remote sensing image-text retrieval," *TGRS*, 2023.

[11] Haojun Jiang, Jianke Zhang, Rui Huang, et al., "Cross-modal adapter for text-video retrieval," *arXiv preprint arXiv:2211.09623*, 2022.

[12] Richang Hong, Daqing Liu, Xiaoyu Mo, et al., "Learning to compose and reason with language tree structures for visual grounding," *TPAMI*, 2019.

[13] Xuyang Liu, Siteng Huang, Yachen Kang, et al., "VGDiffZero: Text-to-image diffusion models can be zero-shot visual grounders," in *ICASSP*, 2024.

[14] Jiabo Ye, Xin Lin, Liang He, et al., "One-stage visual grounding via semantic-aware feature filter," in *ACM MM*, 2021.

[15] Mengyang Sun, Wei Suo, Peng Wang, et al., "A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention," *TMM*, 2022.

[16] Ye Du, Zehua Fu, Qingjie Liu, et al., "Visual grounding with transformers," in *ICME*, 2022.

[17] Ting Liu, Yue Hu, Wansen Wu, et al., "DAP: Domain-aware prompt learning for vision-and-language navigation," in *ICASSP*, 2024.

[18] Yi Xin, Junlong Du, Qiang Wang, et al., "MmAP: Multi-modal alignment prompt for cross-domain multi-task learning," in *AAAI*, 2024.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., "Deep residual learning for image recognition," in *CVPR*, 2016.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, et al., "End-to-end object detection with transformers," in *ECCV*, 2020.