





DARA: Domain- and Relation-aware Adapters Make Parameter-efficient Tuning for Visual Grounding

Ting Liu^{*} Xuyang Liu^{*} Siteng Huang Honggang Chen Quanjun Yin Long Qin Donglin Wang Yue Hu⊠ National University of Defense Technology & Sichuan University & Westlake University







Overview

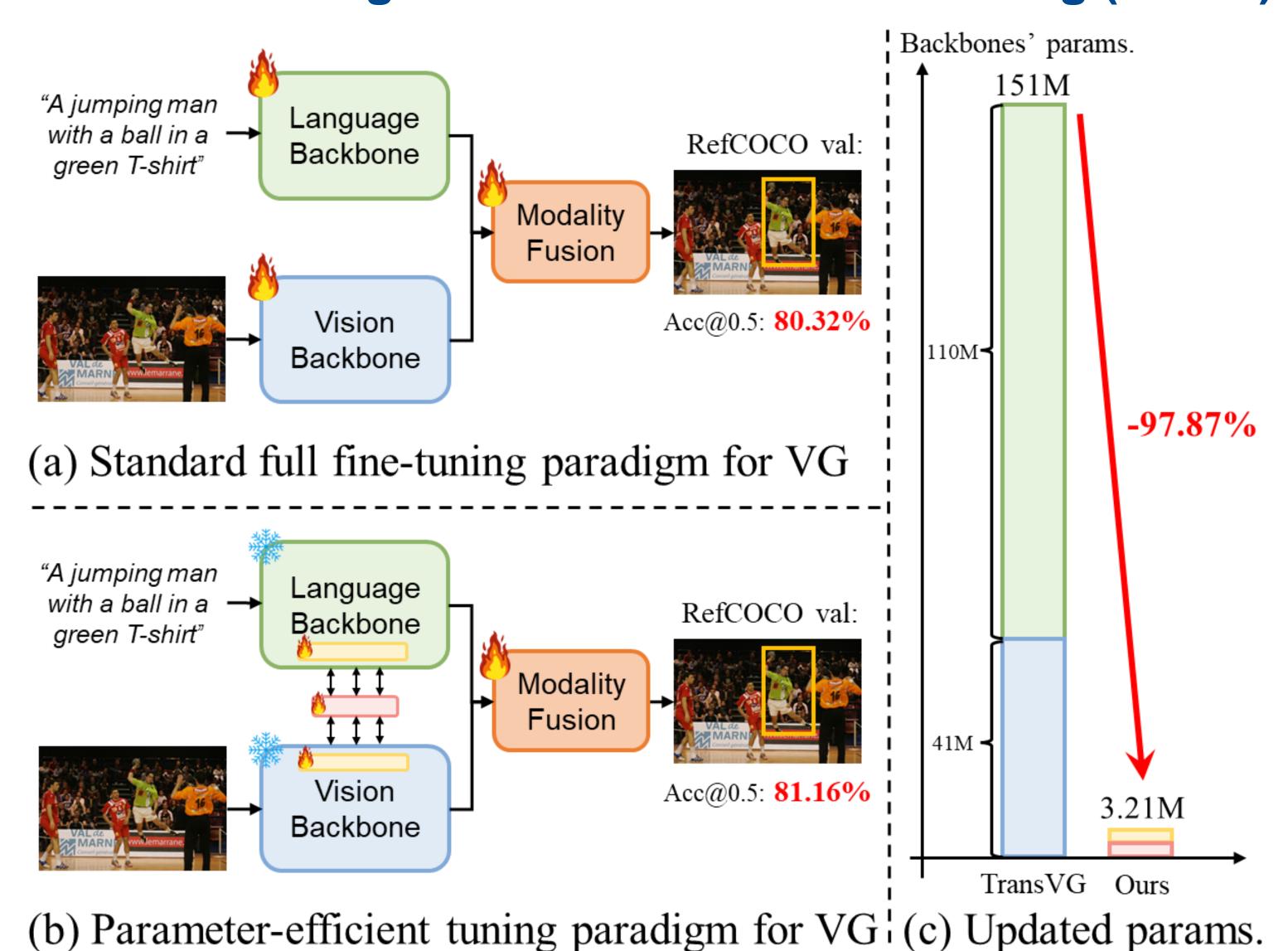
Research Question

How can parameter-efficient transfer learning (PETL) be effectively applied to the visual grounding (VG) to reduce computational costs during fine-tuning, while achieving competitive performance?

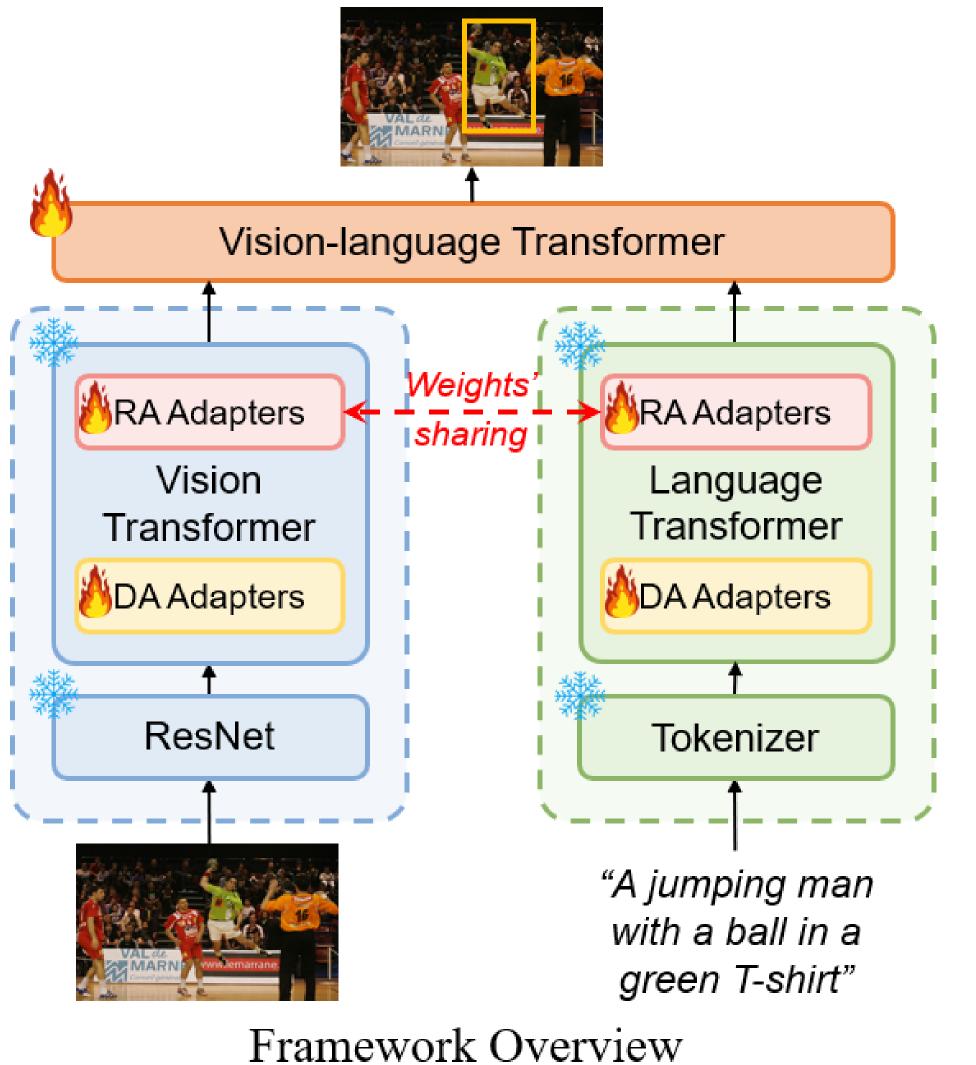
Contributions

- present an in-depth exploration of adapting PETL methods for VG. To the best of our knowledge, this is the first study to investigate parameter-efficient tuning paradigm for visual grounding.
- propose DARA, a novel PETL framework incorporating Domain-aware Adapters and Relation-aware Adapters, to facilitate effective and efficient intra- and inter-modality representation transfer for VG.
- achieve the best accuracy while saving numerous updated parameters compared to full fine-tuning and other PETL methods.

Full Fine-tuning vs Parameter-efficient Tuning (DARA)



Overall Architecture



- Baseline Model: we use a typical transformer-based VG framework, to implement, including: a Vision Backbone, a Language Backbone, and a Vision-language Transformer.
- **DARA:** DARA consists of **Domain**aware Adapters and Relationaware Adapters. During fine-tuning, We freeze the vision and language backbones and update our DARA to facilitate effective and efficient single-modality and cross-modality representations transfer for VG. Notably, with only 2.13% tunable DARA backbone parameters, average accuracy by 0.81% across three benchmarks compared to the baseline model.

Adapter: typically inserting an MLP

Domain-aware Adapters (DA): DA

refine pre-trained intra-modality

representations to be more fine-

grained for the visual grounding

domain, enhancing the precision in

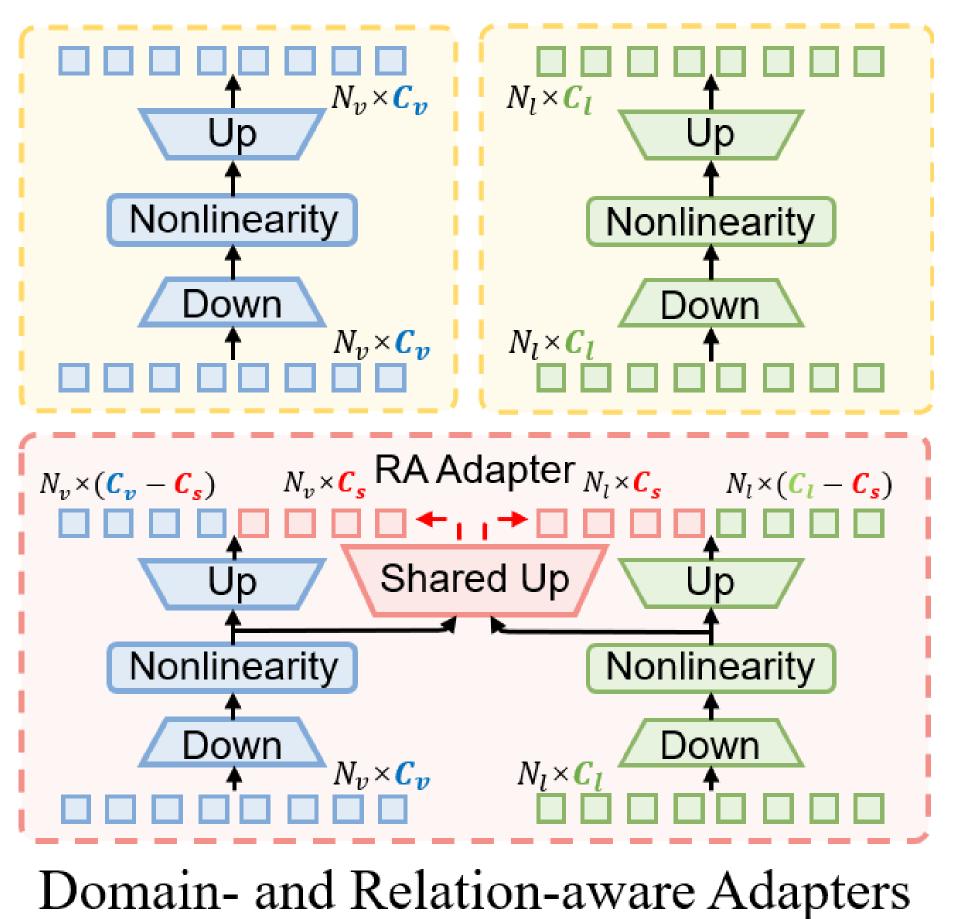
Relation-aware Adapters (RA) :

RA facilitate early cross-modality

describing and locating objects.

into each frozen backbone layer.

Domain-aware and Relation-aware Adapters



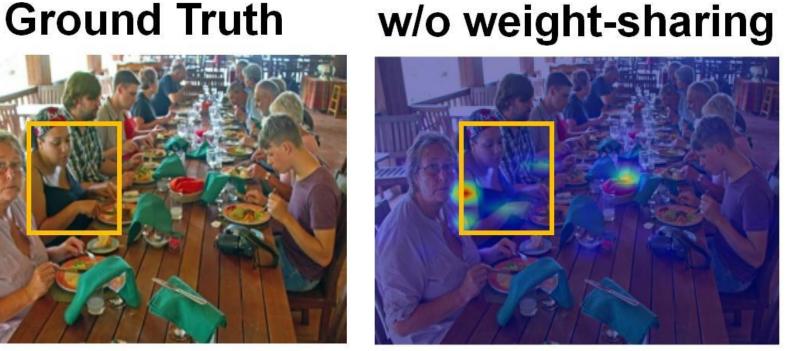
interactions by sharing weights between the vision and language backbones, thus improving spatial reasoning and the effectiveness of fusion for visual grounding.

Quantitative Results

Full fine-tuning methods Farams. Val testA testB val	Methods	Backbone #Updated		RefCOCO			RefCOCO+			RefCOCOg		
Two-stage: MAttNet RN101/LSTM 47M 76.65 81.14 69.99 65.33 71.62 56.02 - 66.58 67.27 RvG-Tree RN101/LSTM 47M 75.06 78.61 69.85 63.51 67.45 56.66 - 66.58 67.27 One-stage: FAOA DN53/LSTM 43M 72.54 74.35 68.50 56.81 60.23 49.60 56.12 61.33 60.26 SAFF DN53/BERT 152M 79.26 81.09 76.55 64.43 68.46 58.43 - 68.94 68.91 PFOS DN53/BERT 152M 77.37 80.43 72.87 63.74 68.54 55.84 61.46 67.08 66.35 Transformer-based: VGTR RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 151M 80.32 82.67 78.12 63	Miculous	(vision/language)	Params.	val	testA	testB	val	testA	testB	val-g	val-u	test-u
MAttNet RvG-Tree RN101/LSTM RN101/LSTM 47M 47M 76.65 78.61 81.14 69.99 65.33 63.51 71.62 67.45 56.02 56.66 - 66.58 66.59 67.27 66.51 One-stage: FAOA DN53/LSTM DN53/BERT 43M 152M 72.54 79.26 74.35 81.09 68.50 72.87 56.81 68.50 60.23 68.43 49.60 56.12 56.12 61.33 60.26 68.94 68.91 68.91 Transformer-based: VGTR RN50/LSTM DN53/BERT 52M 152M 78.70 78.99 82.09 79.71 73.31 72.67 63.57 61.58 66.60 66.60 55.33 62.88 65.62 65.30 66.70 66.03 66.70 66.70 Transformer-based: VGTR RN50/LSTM DN53/BERT 52M 152M 78.70 76.99 82.09 79.71 72.67 72.67 61.58 66.60 66.00 54.00 56.32 66.33 62.88 65.62 65.53 66.70 TransVG† 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 RN50/BERT 3.27M 77.37 78.82 73.37	Full fine-tuning methods											
RvG-Tree RN101/LSTM 47M 75.06 78.61 69.85 63.51 67.45 56.66 - 66.95 66.51 One-stage: FAOA DN53/LSTM 43M 72.54 74.35 68.50 56.81 60.23 49.60 56.12 61.33 60.26 SAFF DN53/BERT 152M 79.26 81.09 76.55 64.43 68.46 58.43 - 68.94 68.91 PFOS DN53/BERT 152M 77.37 80.43 72.87 63.74 68.54 55.84 61.46 67.08 66.35 VGTR RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 151M 80.32 82.67 78.12 63.50 68.15 55.63 66.56 67.66 67.44 TransVG† RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34	Two-stage:									1		
One-stage: FAOA DN53/LSTM 43M 72.54 74.35 68.50 56.81 60.23 49.60 56.12 61.33 60.26 SAFF DN53/BERT 152M 79.26 81.09 76.55 64.43 68.46 58.43 - 68.94 68.91 PFOS DN53/BERT 152M 77.37 80.43 72.87 63.74 68.54 55.84 61.46 67.08 66.35 Transformer-based: VGTR RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70 TransVG¹ RN50/BERT 151M 80.32 82.67 78.12 63.50 68.15 55.63 66.56 67.66 67.44 Parameter-efficient fine-tuming methods Adapter RN50/BERT <td>MAttNet</td> <td>RN101/LSTM</td> <td>47M</td> <td>76.65</td> <td>81.14</td> <td>69.99</td> <td>65.33</td> <td>71.62</td> <td>56.02</td> <td>-</td> <td>66.58</td> <td>67.27</td>	MAttNet	RN101/LSTM	47M	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
FAOA DN53/LSTM 43M 72.54 74.35 68.50 56.81 60.23 49.60 56.12 61.33 60.26 SAFF DN53/BERT 152M 79.26 81.09 76.55 64.43 68.46 58.43 - 68.94 68.91 PFOS DN53/BERT 152M 77.37 80.43 72.87 63.74 68.54 55.84 61.46 67.08 66.35 Transformer-based: VGTR RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70 TransVG† 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 RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	RvG-Tree	RN101/LSTM	47M	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
SAFF DN53/BERT 152M 79.26 81.09 76.55 64.43 68.46 58.43 - 68.94 68.91 PFOS DN53/BERT 152M 77.37 80.43 72.87 63.74 68.54 55.84 61.46 67.08 66.35 **Transformer-based:** VGTR RN50/LSTM DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70 **TransVG† 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 RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 **LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 **AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 **CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 **MRS-Adapter 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	One-stage:			1						1		
PFOS DN53/BERT 152M 77.37 80.43 72.87 63.74 68.54 55.84 61.46 67.08 66.35 Transformer-based: VGTR RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70	FAOA	DN53/LSTM	43M	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.26
Transformer-based: VGTR RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70 TransVG [†] 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 Parameter-efficient fine-tuning methods Adapter RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 <td< td=""><td>SAFF</td><td>DN53/BERT</td><td>152M</td><td>79.26</td><td>81.09</td><td>76.55</td><td>64.43</td><td>68.46</td><td>58.43</td><td>-</td><td>68.94</td><td>68.91</td></td<>	SAFF	DN53/BERT	152M	79.26	81.09	76.55	64.43	68.46	58.43	-	68.94	68.91
VGTŘ RN50/LSTM 52M 78.70 82.09 73.31 63.57 69.65 55.33 62.88 65.62 65.30 DMRNet DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70 TransVG [†] 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 RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	PFOS	DN53/BERT	152M	77.37	80.43	72.87	63.74	68.54	55.84	61.46	67.08	66.35
DMRNet DN53/BERT 152M 76.99 79.71 72.67 61.58 66.60 54.00 - 66.03 66.70 TransVG† 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 RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter RN50/BERT 3.21M 81.16	Transformer-based:											
TransVG† 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 RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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/BE	VGTŘ	RN50/LSTM	52M	78.70	82.09	73.31	63.57	69.65	55.33	62.88	65.62	65.30
Parameter-efficient fine-tuning methods Adapter RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	DMRNet	DN53/BERT	152M	76.99	79.71	72.67	61.58	66.60	54.00	-	66.03	66.70
Adapter RN50/BERT 3.27M 78.02 79.89 75.23 61.35 66.34 54.21 63.18 65.26 66.65 LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	TransVG [†]	RN50/BERT	151M	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44
LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	Parameter-efficient fine-tuning methods											
LoRA RN50/BERT 2.37M 77.57 78.22 73.37 61.24 66.53 53.95 64.27 67.36 66.43 AdaptFormer RN50/BERT 2.38M 76.32 77.16 73.94 60.96 65.19 53.88 61.81 65.44 64.37 CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	Adapter	RN50/BERT	3.27M	78.02	79.89	75.23	61.35	66.34	54.21	63.18	65.26	66.65
CM Adapter RN50/BERT 3.27M 77.37 78.81 74.07 61.34 66.10 53.31 63.93 65.75 64.72 MRS-Adapter 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	·	RN50/BERT	2.37M	77.57	78.22	73.37	61.24	66.53	53.95	64.27	67.36	66.43
MRS-Adapter 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	AdaptFormer	RN50/BERT	2.38M	76.32	77.16	73.94	60.96	65.19	53.88	61.81	65.44	64.37
DARA (ours) RN50/BERT 3.21M 81.16 82.76 76.72 65.58 69.83 57.22 67.21 69.22 67.67	CM Adapter	RN50/BERT	3.27M	77.37	78.81	74.07	61.34	66.10	53.31	63.93	65.75	64.72
A DIFORDING 34364 0.04 0.00 440 3.00 440 0.05 4.50 0.03	MRS-Adapter	RN50/BERT	1.58M	77.14	77.80	74.80	61.13	66.38	53.13	63.07	66.46	65.16
$\Delta_{baseline}$ RN50/BERT 2.13% +0.84 +0.09 -1.40 +2.08 +1.68 +1.59 +0.65 +1.56 +0.23	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

Qualitative Results

Expression 1: "second woman M



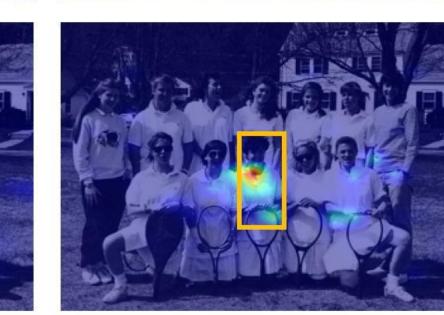


Expression 2: "middle person front row"

on **left**"







Quantitative Results

DARA achieves the best accuracy while ensuring parameter efficiency among all methods, thus validating its effectiveness and efficiency.

Qualitative Results

DARA focuses more on the referred object, which suggests that weightsharing strategy of RA Adapters promotes the synergy between vision and language modalities, thereby enhancing spatial reasoning for VG.

For more experimental results, please refer to our paper.