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

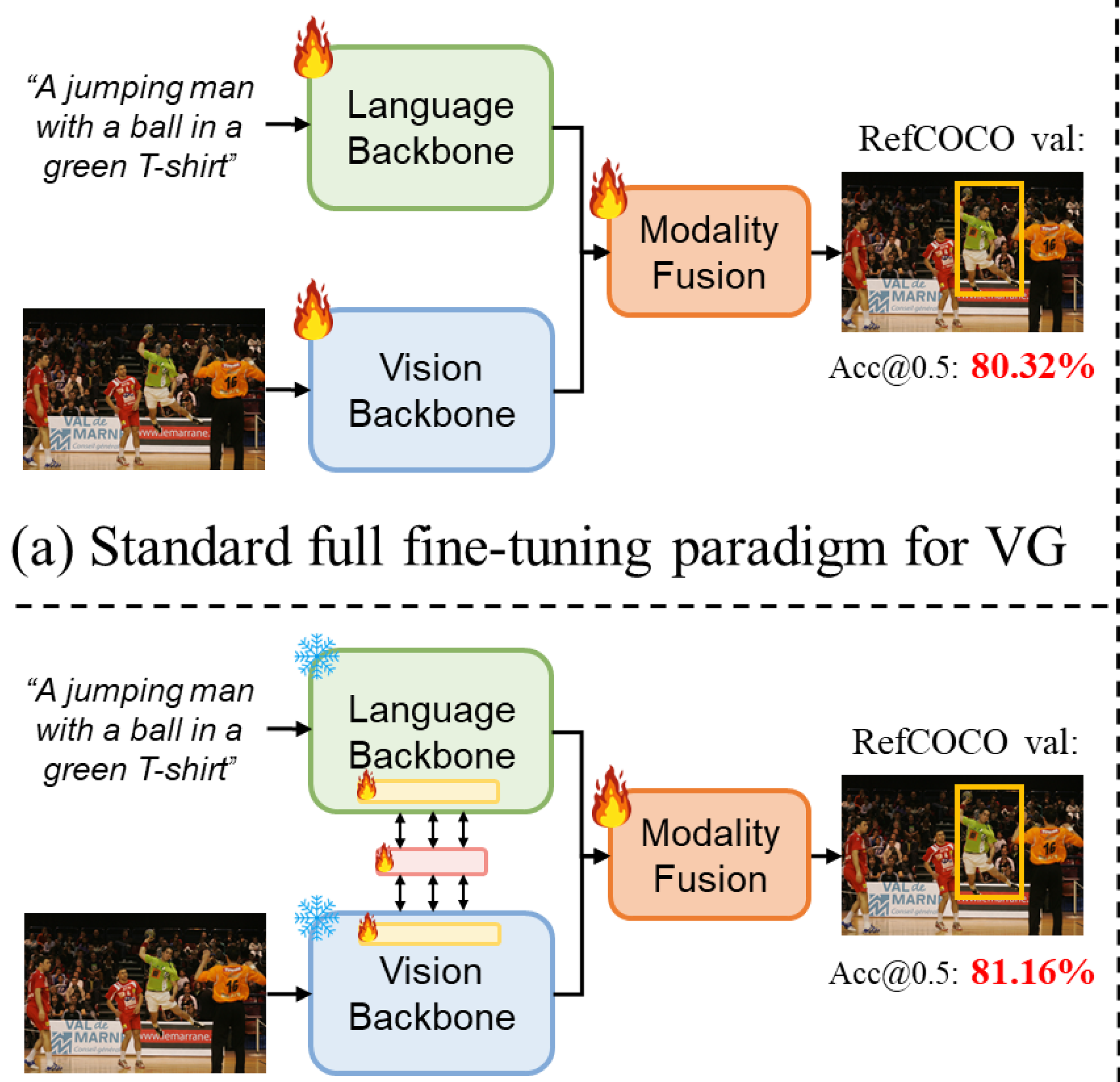
Research Question

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

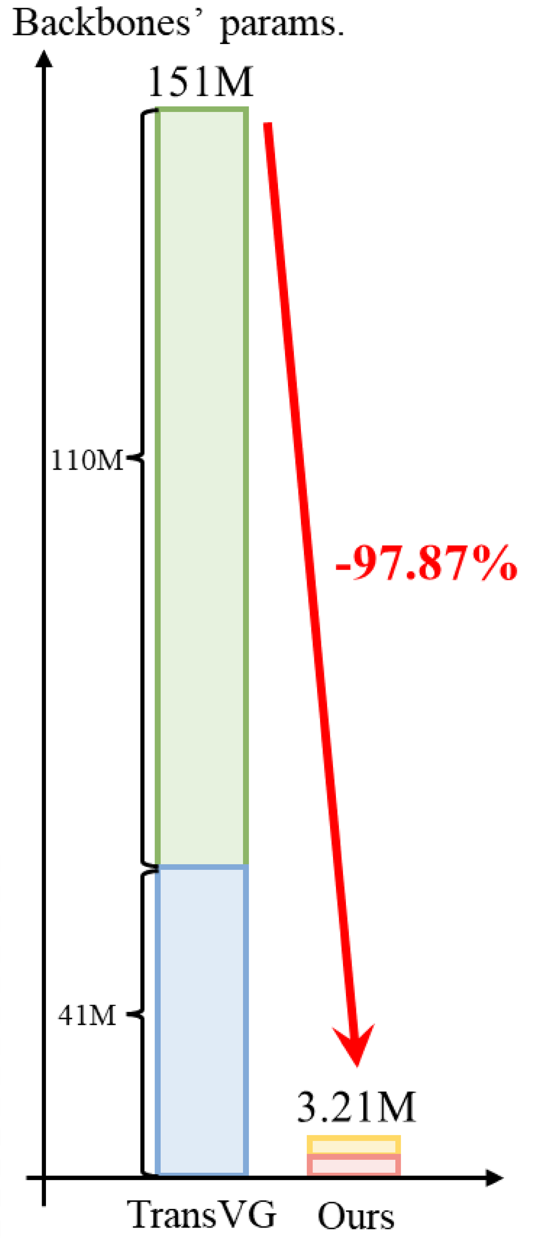
- present an **in-depth exploration** of adapting PETL methods for VG. To the best of our knowledge, this is **the first study** to investigate parameter-efficient tuning paradigm for visual grounding.
- propose DARA, a novel PETL framework incorporating **Domain-aware Adapters** and **Relation-aware 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)



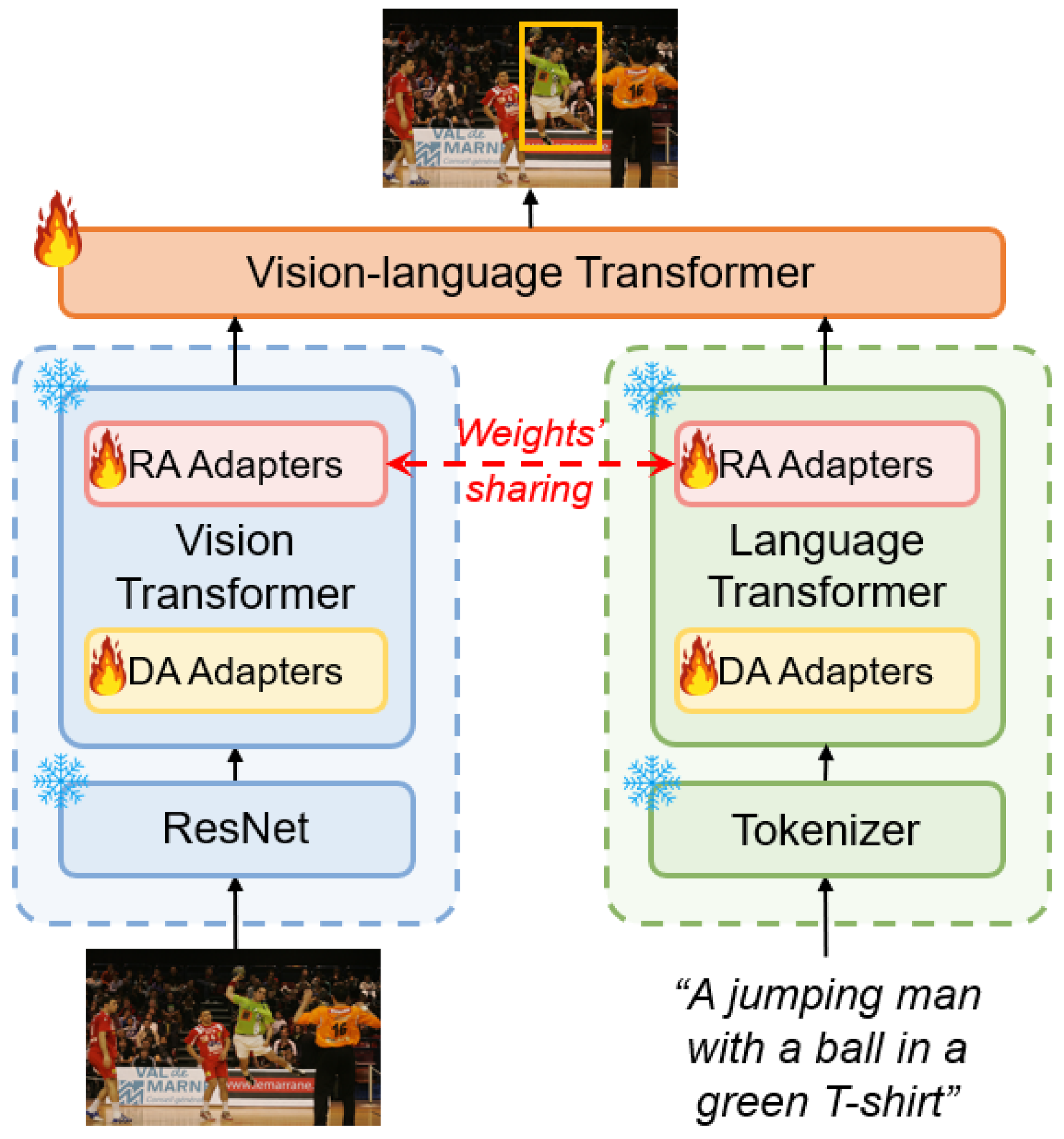
(a) Standard full fine-tuning paradigm for VG

(b) Parameter-efficient tuning paradigm for VG



(c) Updated params.

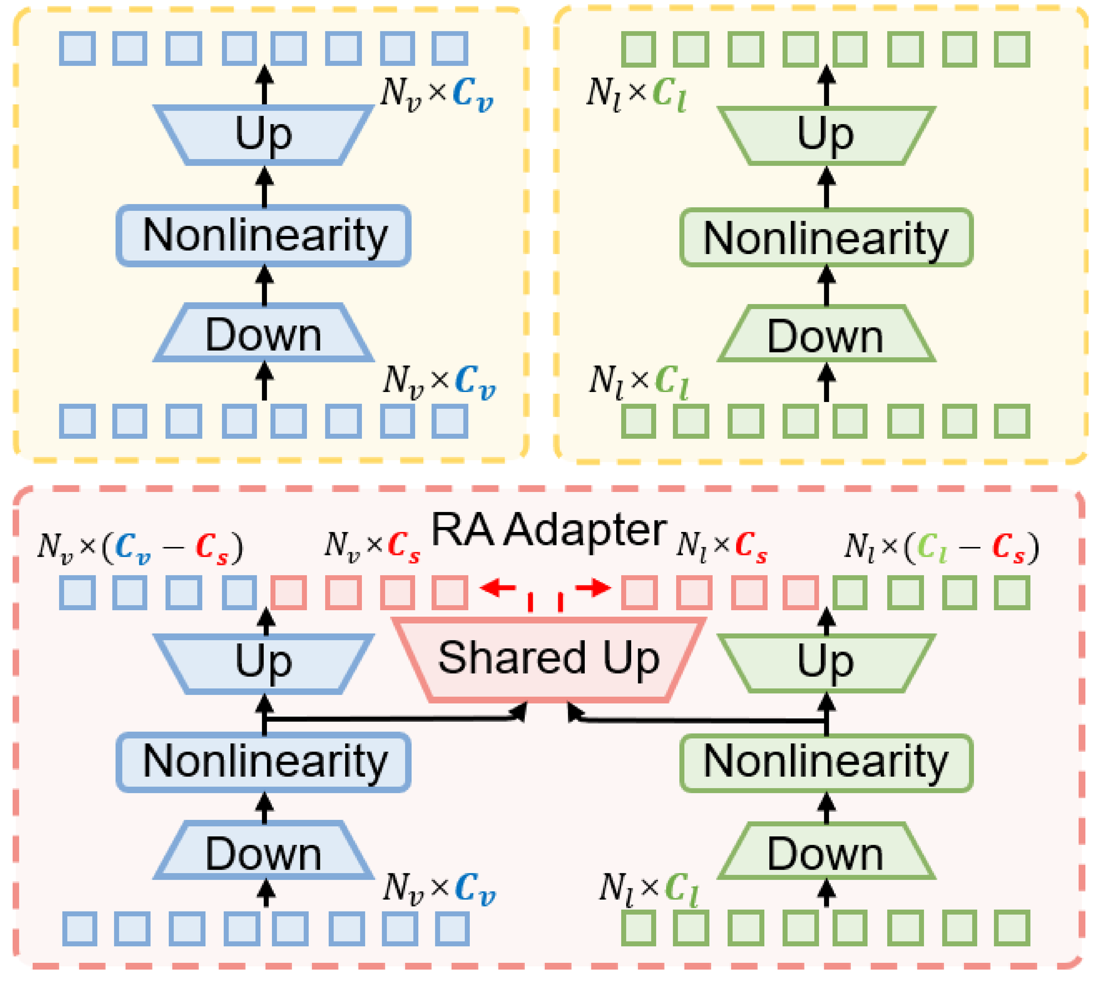
Overall Architecture



Framework Overview

- **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 Relation-aware 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 backbone parameters, DARA improves average accuracy by **0.81%** across three benchmarks compared to the baseline model.

Domain-aware and Relation-aware Adapters



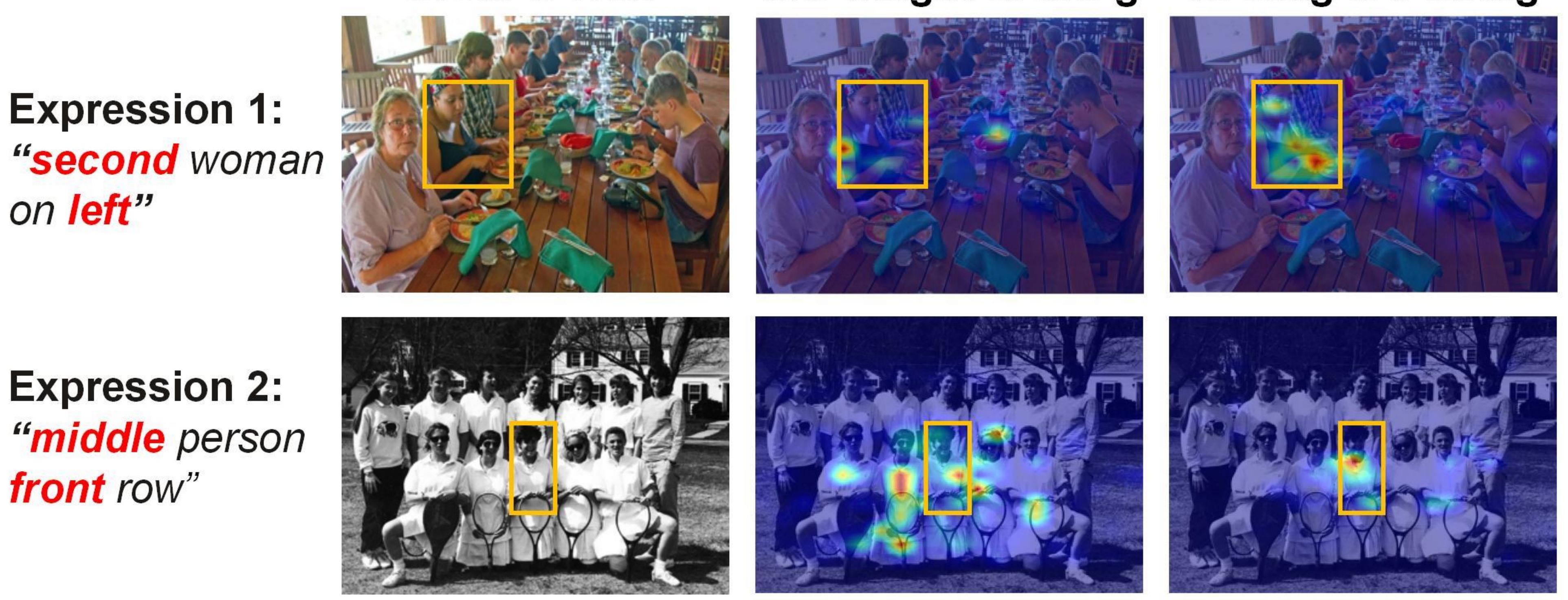
Domain- and Relation-aware Adapters

- **Adapter:** typically inserting an MLP into each frozen backbone layer.
- **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 describing and locating objects.
- **Relation-aware Adapters (RA):** RA facilitate early **cross-modality** interactions by sharing weights between the vision and language backbones, thus improving spatial reasoning and the effectiveness of fusion for visual grounding.

Quantitative Results

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	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	-	66.95	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
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**
$\Delta_{baseline}$	RN50/BERT	2.13%	+0.84	+0.09	-1.40	+2.08	+1.68	+1.59	+0.65	+1.56	+0.23

Qualitative Results



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 weight-sharing 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.