MaPPER: Multimodal Prior-guided Parameter Efficient Tuning for Referring Expression Comprehension

Ting Liu[1]*, Zunnan Xu[2]*, Yue Hu[1], Liangtao Shi[3], Zhiqiang Wang[4], Quanjun Yin[1]†
[1]College of Systems Engineering, National University of Defense Technology,
[2]Tsinghua University, [3]Hefei University of Technology, [4]iFLYTEK Research
liuting20@nudt.edu.cn

Abstract

Referring Expression Comprehension (REC), which aims to ground a local visual region via natural language, is a task that heavily relies on multimodal alignment. Most existing methods utilize powerful pre-trained models to transfer visual/linguistic knowledge by full fine-tuning. However, full fine-tuning the entire backbone not only breaks the rich prior knowledge embedded in the pre-training, but also incurs significant computational costs. Motivated by the recent emergence of Parameter-Efficient Transfer Learning (PETL) methods, we aim to solve the REC task in an effective and efficient manner. Directly applying these PETL methods to the REC task is inappropriate, as they lack the specific-domain abilities for precise local visual perception and visual-language alignment. Therefore, we propose a novel framework of Multimodal Prior-guided Parameter Efficient Tuning, namely MaPPER. Specifically, MaPPER comprises Dynamic Prior Adapters guided by an aligned prior, and Local Convolution Adapters to extract precise local semantics for better visual perception. Moreover, the Prior-Guided Text module is proposed to further utilize the prior for facilitating the cross-modal alignment. Experimental results on three widely-used benchmarks demonstrate that MaPPER achieves the best accuracy compared to the full fine-tuning and other PETL methods with only **1.41%** tunable backbone parameters. Our code is available at https://github.com/liuting20/MaPPER.

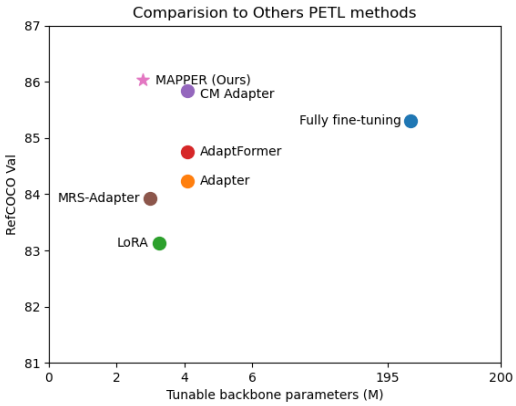

Figure 1: Comparision to others PETL methods.

1 Introduction

Referring Expression Comprehension (REC) (Kamath et al., 2021; Liu et al., 2023; Wu et al., 2023; Bu et al., 2023) is a crucial and challenging task within the multimodal fields, which needs to localize the local image region according to the language expression semantics. REC is fundamental for visual language understanding, with broad applications in fields such as visual-language navigation (Liu et al., 2024a) and human-machine interaction (Chen et al., 2023). Different from vanilla object detection task, REC needs to extract not only global and local spatial information from images, but also relies on the alignment of multimodal features.

Existing approaches (Deng et al., 2021; Kamath et al., 2021; Deng et al., 2023; Shi et al., 2022) transfer the language and vision knowledge from pre-trained models by fully fine-tuning. However, such a fine-tuning strategy is sub-optimal for REC, as reflected in the following aspects: **1)** Fine-tuning the entire backbone might suffer catastrophic forgetting and undermine the extensive prior knowledge learned from pre-training. **2)** The computational cost requirements surge dramatically, particularly for larger foundational models, leading to a significant increase in GPU memory usage. This limits the accessibility of large models for researchers with limited hardware resources.

To address these issues, we shift our focus to Parameter-Efficient Transfer Learning (PETL) (Chowdhury et al., 2023; Wang et al.,

*Equal contribution
†Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4984–4994
November 12-16, 2024

2023a). PETL methods like Adapter tuning and Prompt tuning provide efficient ways to utilize pre-trained models by adjusting a small set of parameters instead of fine-tuning the entire network (Xin et al., 2024c). This approach saves computational resources while still competitive performance improvements. By integrating PETL techniques, we can enhance our models' flexibility and efficiency in adapting to REC. However, we empirically find that directly using these PETL methods cannot achieve satisfactory results in REC (see Figure 1). We argue the main reasons are twofold: **1)** the target objects that require attention in REC often occupy local regions of uncertain size in images, and most existing PETL methods lack the crucial ability to extract multi-scale local semantics for visual perception. **2)** REC is a task that strongly relies on multimodal alignment, and language-oriented adapters are obviously deficient in aligning with visual information. Recently, PETL methods have also been introduced into REC tasks (Xiao et al., 2024; Liu et al., 2024c). HiVG (Xiao et al., 2024) adopts LoRA to fine-tune the frozen CLIP model, but it is not an efficient enough approach due to the heavy alignment design using cross-attention module. In contrast, DARA (Liu et al., 2024c) is a lightweight method in PETL paradigm. However, DARA does not fully address the need for local visual adaptation in the referring expression comprehension, potentially compromising the model's ability to capture fine-grained visual details.

Considering the aforementioned issues, in this paper, we propose a novel framework of **M**ultimod**a**l **P**rior-guided **P**arameter **E**fficient Tuning for **R**EC (MaPPER) that improves text understanding with the aligned prior and enhances vision perception by combining local visual semantics with global perception. As shown in Figure 2, we introduce the vision-aligned text module to generate the aligned prior, which works for the alignment of vision and language feature. Moreover, we insert the Local Convolution Adapter (LoCA) into vision blocks for enhancing visual perception. Specifically, we propose the Dynamic Prior Adapter (DyPA) presented in Figure 3, DyPA can dynamically adjust each token by considering the significance score guided by the aligned prior. In order to promote the interaction of text and vision features, we further propose the Prior-guided Text module (PGT) for fusing the prior and text feature. For the visual branch, most pre-trained visual models are powerful transformer architectures. Unfortunately, vision transformers are observed ignoring local feature details (Peng et al., 2021), which decreases the discriminability between backgrounds and foregrounds. Motivated by this, we introduce the Local Convolution Adapter (LoCA), which integrates multi-scale local knowledge, thereby enhancing the representational power for pre-trained vision transformers. Extensive experiments on RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016; Nagaraja et al., 2016) demonstrate the effectiveness and efficiency of our framework. Our main contributions are as follows:

- We perform an in-depth exploration of parameter-efficient transfer learning (PETL) methods for REC tasks. We introduce MaPPER aimed at improving both the effectiveness and efficiency of visual-text alignment, as well as enhancing visual perception by incorporating local visual semantics.
- We propose the novel Dynamic Prior Adapter (DyPA) and Local Convolution Adapter (LoCA). The former employs aligned prior to dynamically adjust the language encoder, while the latter introduces local visual features for enhancing the visual encoder.
- Extensive experiments demonstrate that our method can outperform the state-of-the-art (SOTA) methods in REC tasks, with only **1.41%** tunable parameters within pre-trained backbones.

2 Related Work

2.1 Referring Expression Comprehension

Referring expression comprehension (REC) (Yu et al., 2018; Yang et al., 2019; Deng et al., 2021; Xiao et al., 2023; Liu et al., 2024e; Xiao et al., 2024) aims to locate a local visual region in images by textual descriptions. Early propose-and-rank methods (Hong et al., 2019; Chen et al., 2019) follow a two-stage pipeline which first utilizes pre-trained object detectors to obtain a set of region proposals, which are then ranked based on their similarity scores with the given textual description. However, these two-stage methods face challenges in terms of the performance of the proposal generators and the additional ranking mechanisms. After the introduction of ViT, the Transformer-based methods (Deng et al., 2021; Du et al., 2022; Yang et al., 2022; Zhu et al., 2022; Liu et al., 2024c; Zhu et al., 2023) have recently emerged that significantly improve the grounding performance. Most

recently, grounding multimodal large language models (Li et al., 2023; Wang et al., 2023b) have propelled the state-of-the-art (SOTA) performance, these works require a large amount of in-domain and other domain datasets. As REC models continue to scale up in size and complexity, fully fine-tuning becomes extremely high training cost.

2.2 Parameter-efficient Transfer Learning

The continuous expansion of pre-trained models demands significant computational resources and consumes considerable storage during fine-tuning (Liu et al., 2024d). To address these challenges, researchers in the NLP and CV domain have explored PETL methods (Hu et al., 2022; Chen et al., 2022; Yuan et al., 2023; Liu et al., 2024b). By focusing on updating only a small subset of parameters, PETL achieves a balance between maintaining high performance and ensuring computational efficiency. This method is particularly advantageous for deploying large-scale models, addressing the challenges posed by increasing model sizes while streamlining the adaptation process to new tasks. The main PETL methods can be classified into three categories: (i) selectively updating a tiny number of existing model parameters (Guo et al., 2020; Zaken et al., 2021); (ii) adjusting newly added parameters to the model or its input (Li and Liang, 2021; Zhou et al., 2022; Xin et al., 2024b); (iii) applying low-rank factorization techniques to the parameters that require updates (Hu et al., 2022; Karimi Mahabadi et al., 2021; Hao et al., 2023; Liu et al., 2024f; Xin et al., 2024a). Some pioneering works like ETRIS (Xu et al., 2023) and DARA (Liu et al., 2024c) sought to utilize adapters to adapt pre-trained models to referring image segmentation and referring expression comprehension, respectively. However, their proposed modules like Bridger (Xu et al., 2023) and RA (Liu et al., 2024c) are insufficient for capturing the complexity of multi-scale local visual features.

3 Methodology

3.1 Framework Overview

The overall framework of the proposed MaPPER is illustrated in Figure 2. Our approach freezes the pre-trained backbone, ensuring parameter efficiency. This framework consists of two distinct efficient tuning modules. The first module, known as the Dynamic Prior Adapter, utilizes aligned prior generated from the Vision-aligned Prior Module to enable efficient modal alignment and adaptation. The second module, referred to as the Local Convolution Adapter module, integrates local visual features into global prior (pre-trained visual knowledge) from the visual encoder, thereby regularizing the whole visual perception. Finally, the complete textual features, alongside aligned prior, are inputted into the Prior-guided Text module for promoting the multimodal alignment.

3.2 Text & Image Feature Extraction

Text Encoder. The REC task relies heavily on word-level understanding due to its concise linguistic expression format, such as "front middle yellow guy", to convey referring information. Owing to its bi-directional encoder representations and the masked language modeling, BERT (Devlin et al., 2018) excels in word-level understanding, making it suitable for text encoding in REC domain. Given the input referring expression T, the text expression is firstly converted into a one-hot vector. Subsequently, each one-hot vector is tokenized into a series of linguistic tokens. A special `[CLS]` token is prefixed to the sequence, and the sequence of tokens is then fed into a stack of 12 transformer encoder layers to progressively capture and model the intricate language tokens.

Visual Encoder. Our work adopts the transformer-based DINOv2-B/14 (Oquab et al., 2023) as the visual backbone. The model involves training the Vision Transformer (ViT) model (Dosovitskiy et al., 2020) on the extensive LVD-142M dataset, utilizing a self-supervised learning strategy. This approach equips the model with the ability to extract powerful visual features, which in turn delivers impressive performance across various downstream tasks. Given an input image $\boldsymbol{I}_0 \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the image is initially divided into N non-overlapping patches, which are then linearly projected into D-dim patch embeddings $\boldsymbol{I}_p \in \mathbb{R}^{N \times D}$. Meanwhile, a learnable `[CLS]` token is prepended to $\boldsymbol{I}_p$, producing $\boldsymbol{I} \in \mathbb{R}^{(N+1) \times D}$.

Considering the substantial number of parameters, we opt to freeze visual and text encoders during the fine-tuning process. This strategy allows for a more efficient allocation of computational resources and focuses the learning on the adjustments of other modules.

3.3 Prior-guided Text Understanding

As detailed in section 3.2, the pre-training mechanism of BERT makes it ideal for the REC task,

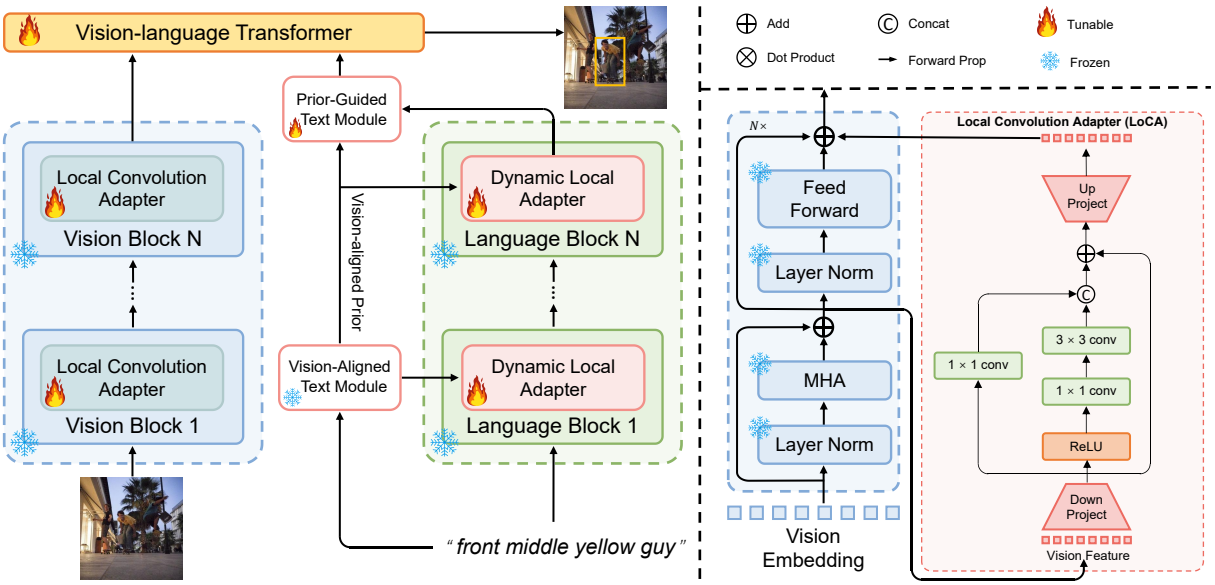

Figure 2: **Overall architecture of MaPPER.** MaPPER freezes the pre-trained vision encoder and language encoder. For the language branch, Dynamic Prior Adapters (DyPA) utilize aligned priors generated from the Vision-aligned Prior Module to enable efficient modal alignment and adaptation. For the language branch, Local Convolution Adapters (LoCA) integrate local visual features the global prior (pre-trained visual knowledge) from the visual encoder. Moreover, the Prior-guided Text module for promoting the multimodal alignment.

which has a relatively high word-level understanding. However, BERT lacks alignment with vision in the pre-training process, and we introduce a Vision-aligned Prior Module to generate a vision-aligned prior. The prior serves for better adjusting BERT encoder, and promoting the interaction of text and vision features.

Vision-aligned Prior Module (VAP). The core of VAP to a produce vision-aligned prior for the REC domain. Considering that CLIP (Radford et al., 2021) model inherently has the ability to align visual with text feature, we used the frozen CLIP followed by a mapping layer M as the VAP module. Given the text input $\boldsymbol{t}$, the vision-aligned prior $\boldsymbol{p}$ can be formulated as follows:

$$\boldsymbol{p} = M(\mathrm{CLIP_f}(\boldsymbol{t})). \tag{1}$$

where the $\mathrm{CLIP_f}$ denotes the frozen CLIP backbone.

Dynamic Prior Adapter (DyPA). To dynamically bridge the gap between the pre-trained BERT model and the complex REC task, we introduce the Dynamic Prior Adapter, which operates in parallel with the text encoder, as shown in Figure 3. DyPA comprising four module: a dynamic scale module (DS), a downward projection with parameters $\boldsymbol{W}^t_{down} \in \mathbb{R}^{r\times d}$, a ReLU activation layer, and an upward projection with parameters $\boldsymbol{W}^t_{up} \in \mathbb{R}^{d\times r}$.

Specifically, we adopt the DS module for integrating the vision-aligned prior $\boldsymbol{p}$ to different layers in the BERT encoder. The module generates scale factors S_f using a scoring weight matrix $\boldsymbol{W}_s \in \mathbb{R}^{1\times d}$, eliminating manual hyper-parameter tuning. Given the prior $\boldsymbol{p}$, the dynamic scaling factor can be formulated as follows:

$$S_f = \mathrm{ReLU}\left(\boldsymbol{p}\boldsymbol{W}_s\right). \tag{2}$$

The downward projection and the upward projection are connected by a ReLU function. In one text encoder layer, the downward projection layer receives processed language tokens $\boldsymbol{x_t}$ from the Multi-head Attention (MHA) layer as input and produces adapted. In general, the output of DyPA $\boldsymbol{x'_t}$ can be described as

$$\boldsymbol{x'_t} = S_f \times \left[\left(\mathrm{ReLU}\left(\boldsymbol{x_t}\boldsymbol{W}^t_{down}\right)\right)\boldsymbol{W}^t_{up}\right]. \tag{3}$$

DyPA utilizes the vision-aligned prior p to dynamically regularize the feed-forward during adapter tuning. To mitigate the influence of Adapter outputs during the initial stages of model training, we initialize $\boldsymbol{W}^t_{up}$ to zero.

Prior-guided Text Module (PGT). Through the design of the DyPA module, we efficiently fine-tune the BERT model to produce fine-grained aligned text features for the REC tasks. In order to promote the interaction of text and vision features for the Multimodal Interactive Module in Sec.3.5, we propose a Prior-Guided Text Module, fusing the prior $\boldsymbol{p} \in \mathbb{R}^{N_t\times C_p}$ into the text features

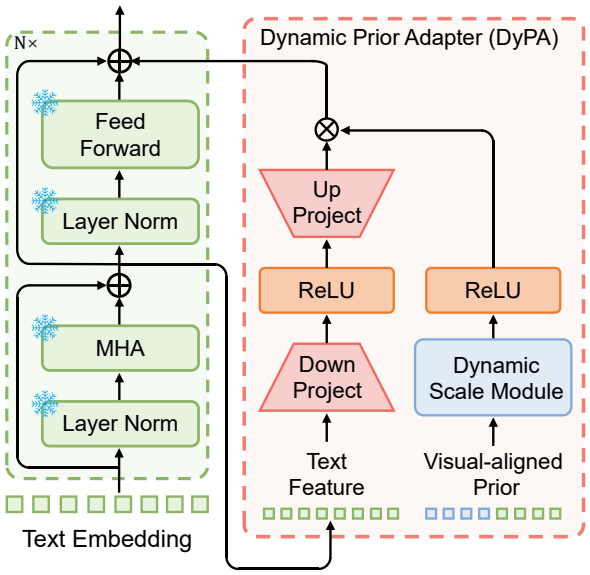

Figure 3: The structure of the Dynamic Prior Adapter.

$\boldsymbol{t} \in \mathbb{R}^{N_t \times C_t}$ generated by the BERT encoder. To achieve this, we employ a projection layer, denoted by $Proj \in \mathbb{R}^{C_p \times C_t}$, to map the prior $\boldsymbol{p}$ onto a transformed representation $\boldsymbol{p'}$. This projection is specifically designed to align the dimensions of the prior with the text features. In order to streamline the process, we concatenate $\boldsymbol{t}$ with the transformed priors $\boldsymbol{p'}$ to get the final text feature $\boldsymbol{f_t}$ integrated with the vision-aligned prior.

$$\begin{aligned} \boldsymbol{p'} &= Proj(\boldsymbol{p'}), \\ \boldsymbol{f_t} &= \text{Concat}[\boldsymbol{p'}, \boldsymbol{t}]. \end{aligned} \tag{4}$$

3.4 Global & Local Visual Perception

For visual perception in the REC task, local features and global representations are important counterparts. Although pre-trained DINOv2 can provide powerful and robust visual features to achieve promising performance, the task-specific visual attention in the REC task often focuses on localized areas of uncertain size in images, which have been visualized in Figure 4.

Local Convolution Adapter (LoCA). To further facilitate the visual perception ability of DINOv2 for the REC task, we propose a Local Convolution Adapter (LoCA) module to adjust the visual foundation models. LoCA introduces the multi-scale local information to further enhance visual perception. The local convolution adapter consists of a down-projection layer $\boldsymbol{W}^v_{down}$, a multi-scale convolution module, a ReLU activation layer, and the up-projection layer $\boldsymbol{W}^v_{up}$.

Specifically, in one visual encoder layer, the downward projection layer receives processed visual tokens $\boldsymbol{x_v}$ from the Multi-head Attention (MHA) layer as input and produces adapted. The multi-scale convolution module consists of two parallel convolutional paths of multi-scale (1×1, 3×3). The 1×1 convolution is strategically placed before the 3×3 convolutions to reduce channel dimension. This design and the bottleneck structure make the local convolution adapter still lightweight. The outputs of the multi-scale convolutional paths are concatenated to form the local feature f_{loc}.

$$\begin{aligned} f_v &= \text{ReLU}(x_v \mathbf{W}^v_{down}), \\ f_{v1} &= \text{Conv}_{1\times1}(f_v), \\ f_{v2} &= \text{Conv}_{3\times3}(\text{Conv}_{1\times1}(f_v)), \\ f_{loc} &= \text{Concat}[f_{v1}, f_{v2}]. \end{aligned} \tag{5}$$

before the up-projection, a skip connection operates in parallel with the multi-scale convolution module.

$$\begin{aligned} f_{loc'} &= f_{loc} + f_v, \\ f_{loc} &= (f_{loc'} \mathbf{W}^v_{up}). \end{aligned} \tag{6}$$

Global and Local Visual Integration. To augment the DINOv2 backbone with multi-scale local visual perception on the REC task, we integrate the Local Convolution Adapter (LoCA) in parallel with the MLP layer within the transformer block. By the concise design, LoCA module adds multi-scale local prior into the DINOv2 model for the REC task. The output of each adapted transformer block can be described as:

$$\begin{aligned} v_l^{mha} &= \text{MHA}(\text{LN}(v_{l-1})) + v_{l-1}, \\ v_l &= \text{MLP}(\text{LN}(v_l^{mha})) + s \cdot f_{loc} + v_l^{mha}. \end{aligned} \tag{7}$$

where s is the scaling factor, and the v_{l-1} represents the previous layer output.

3.5 Multimodal Interactive Module

We have implemented a transformer (Vaswani et al., 2017) architecture that seamlessly integrates multimodal embeddings to forecast the bounding box of the referenced object. Specifically, the adapted vision embeddings $\boldsymbol{f_v} \in \mathbb{R}^{N_v \times C_v}$ and language embeddings $\boldsymbol{f_l} \in \mathbb{R}^{N_l \times C_l}$ are first projected into a common space of joint embeddings $\boldsymbol{f'_v} \in \mathbb{R}^{N_v \times C_p}$ and $\boldsymbol{f'_l} \in \mathbb{R}^{N_l \times C_p}$, both with a unified channel size. Followed by TransVG (Deng et al., 2021) and DARA (Liu et al., 2024c), these joint embeddings,

Methods	Venue	Tuned/Total param.	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	testB	val	testA	testB	val-g	val-u	test-u
Full Fine-tuning											
MAttNet (Yu et al., 2018)	CVPR'18	100%	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
RvG-Tree (Hong et al., 2019)	TPAMI'19	100%	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree (Liu et al., 2019)	ICCV'19	100%	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
FAOA (Yang et al., 2019)	ICCV'19	100%	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.26
ReSC-Large (Yang et al., 2020)	ECCV'20	100%	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
TransVG (Deng et al., 2021)	ICCV'21	100%	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44
QRNet (Ye et al., 2022)	CVPR'22	100%	84.01	85.85	**82.34**	72.94	76.17	63.81	71.89	73.03	72.52
Dynamic-MDETR † (Shi et al., 2022)	TPAMI'23	100%	85.97	88.82	80.12	74.83	**81.70**	63.44	72.21	74.14	74.49
PFOS (Sun et al., 2022)	TMM'22	100%	77.37	80.43	72.87	63.74	68.54	55.84	61.46	67.08	66.35
SeqTR (Zhu et al., 2022)	ECCV'22	100%	81.23	85.00	76.08	68.82	75.37	58.78	-	71.35	71.58
Word2Pix (Zhao et al., 2022)	TNNLS'22	100%	81.20	84.39	78.12	69.46	76.81	61.57	-	70.81	71.34
YORO† (Ho et al., 2023)	ECCV'22	100%	82.90	85.60	77.40	73.50	78.60	64.90	-	73.40	74.30
CLIP-VG (Xiao et al., 2023)	TMM'23	100%	84.29	87.76	78.43	69.55	77.33	57.62	72.64	73.18	72.54
JMRI (Zhu et al., 2023)	TIM'23	100%	82.97	87.30	74.62	71.17	79.82	57.01	69.32	71.96	72.04
MGCross (Miao et al., 2023)	TIP'24	100%	85.10	88.23	80.08	74.44	79.48	65.21	74.50	**77.25**	75.78
TransCP (Tang et al., 2023)	TPAMI'24	100%	84.25	87.38	79.78	73.07	78.05	63.35	72.60	-	-
ScanFormer (Su et al., 2024)	CVPR'24	100%	83.40	85.86	78.81	72.96	77.57	62.50	74.10	-	74.14
Parameter-efficient Transfer Learning											
DARA (Liu et al., 2024c)	Arxiv'24	1.63%	81.16	82.76	76.72	65.58	69.83	57.22	67.21	69.22	67.67
MaPPER (Ours)	-	**1.41%**	**86.03**	**88.90**	81.19	**74.92**	81.12	**65.68**	**74.60**	76.32	**75.81**

Table 1: **Comparison with latest SOTA methods on RefCOCO/+/g for visual grounding. † indicates that all of the RefCOCO/+/g training data has been used during pre-training.** "Tuned/Total param." is the average percentage of tuned parameters in backbone. We highlight the **best** and the second-best results.

along with a learnable [REG] token, are processed through a series of six transformer encoder layers, to fuse the cross-modality embeddings. Finally, a prediction head, implemented as a Multi-layer Perceptron with two 256-dimensional hidden layers and a linear output layer, takes the [REG] token as input and projects it onto the 4-dimensional coordinates for defining the bounding box.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We validate our method on three widely-used REC benchmarks: RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016; Nagaraja et al., 2016). We follow the previous research that employs top-1 accuracy (%) as the evaluation metric. Specifically, a prediction is deemed accurate only when its IoU exceeds or equals 0.5. In addition to Precision@0.5, we also report the number of tunable parameters in the pre-trained encoders to compare the fine-tuning efficiency with traditional full fine-tuning and other PETL methods.

Implementation Details. The vision encoder is initialized with DINOv2-B/14 (Oquab et al., 2023), while the language encoder uses BERT-base (Devlin et al., 2018). The resolution of the input image is 518×518. Both the DINOv2-B/14 model and the BERT-base model process tokens with a feature dimension of 768. The Multimodal Interactive Module uses Xavier initialization. DyPA are initialized with Kaiming normal initialization and inserted into the transformer layers for the language encoder. The bottleneck dimension C_d for DyPA is 32. For LoCA, the 1×1 convolution before the 3×3 convolution reduces the channel to 24. The output dimensions of the two convolutional paths are 192 and 96, so the input dimension of the these convolutional paths is 288. For fair comparisons, PETL methods in Table 2 use the same base architecture, and keeping the vision and language encoder fixed.

4.2 Main Results

We conducted a comprehensive comparison between our proposed MaPPER model and a series of previous referring expression comprehension (REC) methods. The main experimental results are presented in Table 1, from which we can observe that: MaPPER achieves the best accuracy while ensuring parameter efficiency among all methods, thus validating its effectiveness and efficiency.

Effectiveness. As Table 1 shown, on the three commonly challenging benchmarks, MaPPER outperforms all traditional full fine-tuning methods. Compared to DARA (Liu et al., 2024c), a parameter-efficient transfer learning method, we achieves best results on the three benchmarks. Notably, even compared to some methods that are pre-trained on

Methods	Params.↓ (M)	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val-g	val-u	test-u
Fully fine-tuning	196	85.31	87.80	81.03	74.57	80.22	65.31	73.76	74.22	75.02
Adapter (Houlsby et al., 2019)	4.09	84.23	86.76	79.98	73.76	79.91	65.14	72.37	74.19	74.25
LoRA (Hu et al., 2022)	3.25	83.13	85.51	78.32	73.66	78.73	64.85	73.67	74.66	74.83
AdaptFormer (Chen et al., 2022)	4.09	84.75	86.14	79.73	73.05	79.63	65.26	72.19	73.93	74.36
CM Adapter (Jiang et al., 2022)	4.09	85.84	86.49	79.67	74.06	79.91	64.27	73.61	73.54	74.19
MRS-Adapter (Yuan et al., 2023)	2.98	83.92	85.06	78.52	71.13	78.38	63.13	72.42	73.26	72.92
MaPPER	**2.77**	**86.03**	**88.90**	**81.19**	**74.92**	**81.12**	**65.68**	**74.60**	**76.32**	**75.81**

Table 2: **Comparison with PETL methods using the DINO-B Backbone on RefCOCO, RefCOCO+ and RefCOCOg.** "Param." indicates the number of tunable parameters in the pre-trained encoders. To ensure fairness, we kept the original parameter settings from previous methods.

#	Local Conv. Adapter	Params. (M)	RefCOCO		
			val	testA	testB
(a)		0	82.37	84.13	77.57
(b)	✓	1.58	84.28	86.02	79.38

Table 3: **Effectiveness of Local Convolution Adapter (LoCA)** for the visual branch. Note the ablation study without adding any component in the text branch, and we freeze the text encoder. (a) represents freezing both the text and visual branche.

#	Multi-scale size	Params. (M)	RefCOCO		
			val	testA	testB
(a)	1×1	1.48	83.51	85.35	78.56
(b)	1×1 3×3	1.58	84.28	86.02	79.38
(c)	1×1 3×3 5×5	1.70	83.98	85.72	79.02

Table 4: **Effectiveness of multi-scale size** for the visual branch.

the the RefCOCO/+/g (indicated by † in Table 1), our MaPPER model achieves the highest scores across all evaluation tasks, with particularly strong performance on the RefCOCO+, which present greater challenges compared to RefCOCO.

Efficiency. Table 1 clearly illustrates that MaPPER not only achieves the best performance, but also highlights its huge advantages in parameter efficiency. MaPPER reduced the tunable backbone parameters by 98.59% compared to the traditional full fine tuning method. Compared to the PETL method DARA (Liu et al., 2024c), MaPPER has also lower tunable parameters.

4.3 Comparison with Other PETL Methods

We conduct experiments comparing our MaPPER with other parameter-efficient tuning methods using DINOv2-Base as the backbone. To ensure fairness, we retain the original parameter settings from previous methods and adjust the bottleneck to achieve comparable parameter counts. Table 2 illustrates that MaPPER outperforms other PETL methods on all three benchmarks, and even performs better than fully fine-tuning. This highlights the effectiveness of MaPPER in adapting pre-trained knowledge for the REC domain. Through introducing vision-aligned prior, MaPPER enhance the modeling of the vision-text alignment capability. Furthermore, inserting Local Convolution Adapters into DINOv2, making it more suitable for REC tasks with enhanced local visual perception. Previous PETL methods lack these functionalities, rendering them less effective for REC tasks. To summarize, by the specific design for the REC domain, MaPPER achieves superior performance with only **2.77** million parameters.

4.4 Ablation Study

Effectiveness of Local Convolution Adapter. We assess the impact of the Local Convolution Adapter (LoCA) by performing an ablation study and reporting the results on RefCOCO validation and test datasets. From Table 3, it is evident that introducing the LoCA yields a great improvement, increasing the average performance to 1.87%. This indicates that the LoCA enhances the visual perception of DINOv2 with local visual feature.

Effect of Multi-scale Size for Visual Branch. To further verify the effect of local visual information, we perform the attempts of using only a single-size convolution kernel (1×1), and three scales (1×1 3×3 5×5). Table 4 indicates that it is difficult for an adapter with a single-size convolution kernel (a) to perform well for the REC. Local Features are too fine-grained (c) are also not optimal. In contrast, appropriate multi-scale (b) provide proper local information, thus achieving the best performance.

Effect of the Vision-aligned Prior for Text Branch. From Table 5, we can see that: **(1)** Freezing the text encoder while only tuning local convolution adapter can also brings great perfor-

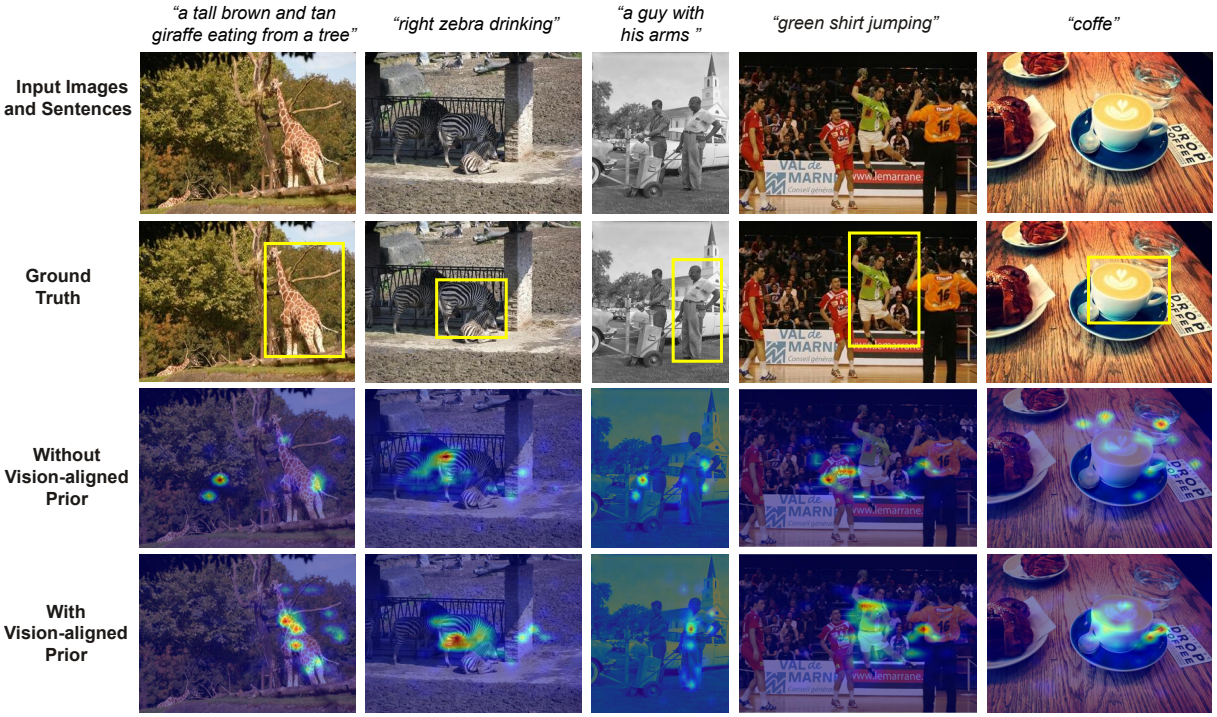

Figure 4: Visualizations of attention maps from the Multimodal Interactive Module for validating the effect of the vision-aligned prior.

#	Adapter w/o $\boldsymbol{p}$	Adapter w. $\boldsymbol{p}$ (DyPA)	Params. (M)	RefCOCO val	testA	testB
(a)			1.58	84.28	86.02	79.38
(b)	✓		1.73(+0.15)	84.78	86.62	79.89
(c)		✓	1.79(+0.21)	85.32	87.62	80.12

#	Adapter w. $\boldsymbol{p}$ (DyPA)	PGT	Params. (M)	RefCOCO val	testA	testB
(d)		✓	2.56(+0.98)	85.76	87.88	80.98
(f)	✓	✓	2.77(+1.19)	**86.03**	**88.90**	**81.19**

Table 5: **Effectiveness of the Vision-Prior for the text branch**. Note (a) represents freezing the text encoder while tuning the LoCA in the visual encoder, and the **LoCA included in Params.**. (b) represents using the vanilla adapter without $\boldsymbol{p}$. (d) represents only using the PGT without any adapters.

mance (Table 5 (a)); **(2)** it is crucial to obtain a dynamic scale with the vision-prior, the Dynamic Prior Adapter (DyPA) brings better performance compared to vanilla adapter fixing the scale to 1.0 (Table 5 (b,c)); **(3)** by the design of Prior-guided Text Module (PGT), we further promote the interaction of text and vision features (Table 5 (f)); **(4)** Incorporating the DyPA and PGT results in an average improvement of 1.02% compared to only using DyPA.

4.5 Qualitative Results

To investigate the impact of vision-aligned prior, we visualize the attention maps from the Multi-modal Interactive Module under two strategies: with and without the vision-aligned prior. In the absence of the prior represents the text adapter without dynamic scale, and the prior-guided text module is not introduced. As shown in Fig. 4, referring expressions contain object appearance attributions, human actions, and spatial relationships. It is observable that the model can focus well on the local target region of the whole image with the vision-aligned prior. This indicates that vision-aligned prior enhancing the alignment ability of MaPPER.

5 Conclusion

In this study, we present an innovative Parameter-Efficient Transfer Learning (PETL) approach designed for multi-modal language grounding tasks, especially in referring expression comprehension. MaPPER enhances the adapters with multi-modal prior through the implementation of a simple yet effective fine-tuning strategy. We aims at improving both the effectiveness and efficiency of visual-text alignment, as well as enhancing visual perception by incorporating local visual semantics. The Dynamic Prior Adapter (DyPA) employs aligned prior to dynamically adjust the language encoder, while the Local Convolution Adapter (LoCA) introduces local visual features for enhancing the visual encoder. MaPPER not only surpasses the performance of fully fine-tuned models but also

does more efficiently.

6 Limitation

While our proposed method has shown enhanced efficiency, scalability, and parameter optimization in the realm of REC tasks, surpassing conventional fully fine-tuned models, our empirical inquiries have been confined to this specific domain. It is imperative for future research to broaden the validation scope encompass include variety range of other multi-modal tasks. Moreover, while our approach can effectively decrease the quantity of parameters necessitating training, thus conserving computational and storage resources, it still mandates a training process. As the frontier of multi-modal large-scale models progresses, there is a significant opportunity for future exploration into open-vocabulary zero-shot referring expression comprehension. This area of research could unveil innovative pathways and contribute to the evolution of models capable of comprehending and generating expressions without the constraint of prior training.

References

Yuqi Bu, Xin Wu, Liuwu Li, Yi Cai, Qiong Liu, and Qingbao Huang. 2023. Segment-level and category-oriented network for knowledge-based referring expression comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Proceedings of the Advances in Neural Information Processing Systems*.

Yi Wen Chen, Yi Hsuan Tsai, Tiantian Wang, Yen Yu Lin, and Ming Hsuan Yang. 2019. Referring expression object segmentation with caption-aware consistency. In *Proceedings of the British Machine Vision Conference*.

Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. 2023. APoLLo: Unified adapter and prompt learning for vision language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. 2023. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2022. Visual grounding with transformers. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Demi Guo, Alexander M Rush, and Yoon Kim. 2020. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*.

Tianxiang Hao, Hui Chen, Yuchen Guo, and Guiguang Ding. 2023. Consolidator: Mergable adapter with group connections for vision transformer. In *International Conference on Learning Representations*.

Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. 2023. Yorolightweight end to end visual grounding. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 3–23. Springer.

Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. 2019. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. 2022. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035.

Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. 2023. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *International Conference on Computer Vision (ICCV), Oral*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1728–1736.

Ting Liu, Yue Hu, Wansen Wu, Youkai Wang, Kai Xu, and Quanjun Yin. 2024a. Dap: Domain-aware prompt learning for vision-and-language navigation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Ting Liu, Yue Hu, Wansen Wu, Youkai Wang, Kai Xu, and Quanjun Yin. 2024b. Panda: Prompt-based context-and indoor-aware pretraining for vision and language navigation. In *MultiMedia Modeling: 30th International Conference*.

Ting Liu, Xuyang Liu, Siteng Huang, Honggang Chen, Quanjun Yin, Long Qin, Donglin Wang, and Yue Hu. 2024c. DARA: Domain- and relation-aware adapters make parameter-efficient tuning for visual grounding. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Ting Liu, Xuyang Liu, Liangtao Shi, Zunnan Xu, Siteng Huang, Yi Xin, and Quanjun Yin. 2024d. Sparse-Tuning: Adapting vision transformers with efficient fine-tuning and inference. *arXiv preprint arXiv:2405.14700*.

Xuyang Liu, Siteng Huang, Yachen Kang, Honggang Chen, and Donglin Wang. 2024e. VGDiffZero: Text-to-image diffusion models can be zero-shot visual grounders. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Xuyang Liu, Ting Liu, Siteng Huang, Yue Hu, Quanjun Yin, Donglin Wang, and Honggang Chen. 2024f. M^2ist: Multi-modal interactive side-tuning for memory-efficient referring expression comprehension. *arXiv preprint arXiv:2407.01131*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Peihan Miao, Wei Su, Gaoang Wang, Xuewei Li, and Xi Li. 2023. Self-paced multi-grained cross-modal interaction modeling for referring expression comprehension. *IEEE Transactions on Image Processing*.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. 2021. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. 2022. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. 2024. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13449–13458.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*.

Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. 2023. Context disentangling and prototype inheriting for robust visual grounding. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*.

Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu, et al. 2023a. APrompt: Attention prompt tuning for efficient adaptation of pre-trained language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Cantao Wu, Yi Cai, Liuwu Li, and Jiexin Wang. 2023. Scene graph enhanced pseudo-labeling for referring expression comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. 2024. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. *arXiv preprint arXiv:2404.13400*.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. 2023. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*.

Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. 2024a. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16085–16093.

Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. 2024b. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16076–16084.

Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. 2024c. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*.

Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. 2023. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Proceedings of the European Conference on Computer Vision*.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. 2022. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15502–15512.

Licheng Yu, Zhe Lin, Xiaohui Shen, et al. 2018. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*.

Yuan Yuan, Yang Zhan, and Zhitong Xiong. 2023. Parameter-efficient transfer learning for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Heng Zhao, Joey Tianyi Zhou, and Yew-Soon Ong. 2022. Word2pix: Word to pixel cross-attention transformer in visual grounding. *IEEE Transactions on Neural Networks and Learning Systems*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. SeqTR: A simple yet universal network for visual grounding. In *Proceedings of the European Conference on Computer Vision*.

Hong Zhu, Qingyang Lu, Lei Xue, Mogen Xue, Guanglin Yuan, and Bineng Zhong. 2023. Visual grounding with joint multi-modal representation and interaction. *IEEE Transactions on Instrumentation and Measurement*.