

SwimVG: Step-wise Multimodal Fusion and Adaption for Visual Grounding

Liangtao Shi, Ting Liu, Xiantao Hu, Yue Hu, Quanjun Yin, Richang Hong, Senior Member, IEEE

Abstract— Visual grounding aims to ground an image region through natural language, which heavily relies on cross-modal alignment. Most existing methods transfer visual/linguistic knowledge separately by fully fine-tuning uni-modal pre-trained models, followed by a simple stack of visual-language transformers for multimodal fusion. However, these approaches not only limit adequate interaction between visual and linguistic contexts, but also incur significant computational costs. Therefore, to address these issues, we explore a step-wise multimodal fusion and adaption framework, namely SwimVG. Specifically, SwimVG proposes step-wise multimodal prompts (Swip) and cross-modal interactive adapters (CIA) for visual grounding, replacing the cumbersome transformer stacks for multimodal fusion. Swip can improve the alignment between the vision and language representations step by step, in a token-level fusion manner. In addition, weight-level CIA further promotes multimodal fusion by cross-modal interaction. Swip and CIA are both parameter-efficient paradigms, and they fuse the cross-modal features from shallow to deep layers gradually. Experimental results on four widely-used benchmarks demonstrate that SwimVG achieves remarkable abilities and considerable benefits in terms of efficiency. Our code is available at <https://github.com/liuting20/SwimVG>.

Index Terms—vision and language, multimodal representation, visual grounding.

I. INTRODUCTION

Visual grounding (VG) [1]–[4] refers to locating the bounding box region described by a textual expression in a specific image, which is one of the most challenging tasks in multimodal fields. In contrast to vanilla detection tasks, VG requires fine-grained vision-language alignment so as to precisely locate an object described through a language expression. The evolution of VG has considerable potential to promote vision-language understanding, and enjoys broad applications in fields such as robot navigation [5], visual Q&A [6] and automatic driving [7], [8].

* Liangtao Shi and Ting Liu contributed equally to this paper.

† Richang Hong is the corresponding author of this paper.

Liangtao Shi and Richang Hong with the Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Hefei 230009, China, and also with the Ministry of Education and School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: shilt@mail.hfut.edu.cn; hongrc.hfut@gmail.com).

Ting Liu, Yue Hu and Quanjun Yin are with School of systems engineering, National University of Defense Technology, Changsha, Hunan Province, 410073, China. (e-mail: liuting20@nudt.edu.cn; yquanjun@126.com; huyue11@nudt.edu.cn). Xiantao Hu with the Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210014, China (e-mail: huxiantao481@gmail.com). This research was partially supported by the National Natural Science Fund of China (Grant Nos. 62306329 and 62103425), Natural Science Fund of Hunan Province (Grant Nos. 2023JJ40676 and 2022JJ40559).

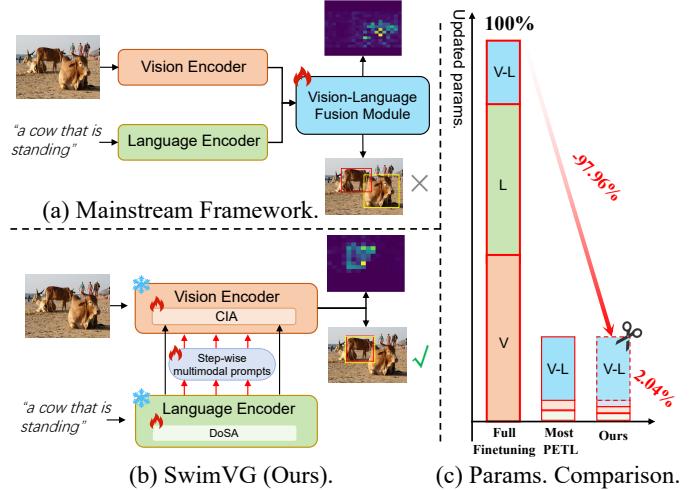


Fig. 1. Comparison of multimodal fusion strategy between (a) mainstream framework and (b) SwimVG (ours) for visual grounding. Freezing the pre-trained models (❄️) and only updating (🔥) the tiny modules in SwimVG reduces 97.96% updated parameters while achieving even stronger performance.

Early visual grounding methods followed target detection frameworks, evolving from the initial two-stage approaches to the recent one-stage methods. Benefiting from the open source of transformer-based pre-trained models, a growing number of approaches [1], [9]–[11] transfer the language and vision knowledge from pre-trained models by fully fine-tuning, such as TransVG [9], TransVG++ [10], and VG-LAW [2]. These methods commonly adopt visual and textual encoders to extract features, respectively, which are subsequently input into a vision-language (VL) transformer for cross-modal interaction. As shown in Fig.1(a), we visualize the last layer of vision-language transformer in mainstream method, it indicates that the visual attentions focus on the foreground area of the image, rather than the text-relevant region (“standing”). We have summarized two reasons for this phenomenon:

- The vision-language encoder for multimodal fusion is a coarse stack of transformers, the mechanism not only limits the sufficient interaction between vision and language contexts, but also exacerbates the computational cost due to the deep transformer-based structure.
- Fine-tuning the entire backbone might suffer catastrophic forgetting and undermine the extensive prior knowledge learned from pre-training. In addition, fully training large pre-trained models and the VL transformer can be computationally expensive and time-consuming in practice.

Several previous works have noticed the insufficient interaction problem, QRNet [12] achieves the expression-aware visual feature extraction by inserting carefully designed interaction modules into the visual backbone. VG-LAW [2] proposes a language adaptive weight generator, generating weights for the fusion of visual and text features. However, they still require fully fine-tuning backbone and the sophisticated designs of interactive modules. More recently, Parameter-Efficient Transfer Learning (PETL) methods have also been introduced into visual grounding [13], [14], HiVG [13] adopts LoRA to fine-tune the frozen CLIP model, and DARA [14] designs DA adapter and RA adapter to transfer intra- and inter-modality representations for the VG domain. However, due to the simple fusion strategy of vision-language transformer, they are not sufficient for multimodal alignment, which could potentially compromise the model’s ability to capture text-relevant visual details.

In this paper, we aim to explore an efficient tuning and lightweight cross-modal interaction strategy. Inspired by the efficiency of Prompt Tuning [15]–[17] and Adapter [18], which only require fine-tuning a tiny number of parameters to adapt pre-trained models to various downstream tasks. We propose a step-wise multimodal fusion and adaption framework (SwimVG). As depicted in Fig. 1(b), we design step-wise multimodal prompts (Swip) for multimodal fusion step by step, and explore a cross-modal interactive adapter (CIA) for further vision-text alignment. The visualizations of Swip (Fig. 7(b)) and CIA (Fig. 7(c)) demonstrate that both of them can independently facilitate multimodal interaction. Their integration, namely SwimVG, as visualized in Fig. 7(c), leads to enhanced multimodal fusion. Through these elaborate designs, we implement an efficient and effective multimodal fusion strategy, abandoning the additional vision-language transformers used in previous methods [9], [10], [13], [14], [19]. As shown in Fig. 2 (b), the vision attentions of the last layer in vision encoder indicate that SwimVG focuses exactly the text-relevant region.

Specifically, to efficient tuning the whole network, we frozen the vision and text backbone, and adopt domain-specific adapters (DoSA) for transferring pre-trained language knowledge to the specific task. To achieve adequate multimodal alignment, we investigate two strategies, namely token-level and weight-level. For token-level multimodal fusion, we design step-wise multimodal prompts, which is formed by gradually integrating a learnable token that can represent the global text semantics into the visual backbone layer by layer. These tokens are initially placed on the language encoder layers, and then mapped to the visual encoder from shallow to deep layers. To further enhance the multimodal fusion in a weight-level manner, we propose a novel cross-modal interactive adapter, which integrate visual and textual features by multi-head cross-attention mechanism. The multimodal adaptation process involves a set of low-rank weight matrices reorganized, producing the crucial alignment capabilities for visual grounding. By the multi-level design of token- and weight-level, for a given image input, the visual encoder can focus more on the text-relevant area, without fully fine-tuning the pre-trained models.

We conduct extensive experiments on RefCOCO [20], RefCOCO+ [20], RefCOCOg [21], [22] and Flickr30K Entities [23], and our method achieves state-of-the-art (SOTA) performance on the four widely used datasets. In addition, we demonstrate the efficiency of our framework in Table III, it can be seen that the inference time of SwimVG is about 40% faster than these mainstream methods using the vision-language transformer. The main contributions can be summarized as three-fold:

- We proposed a concise and efficient framework of multi-modal fusion and adaption, which adapt pre-trained models to visual grounding step by step. SwimVG achieves token-level and weight-level interaction between visual and language representations, and significantly alleviates the task gap between pre-trained models and grounding.
- We replace the heavyweight vision-language transformer with cross-modal interactive adapters and step-wise multimodal prompts, which allow for fine-tuning the entire model in a lightweight manner.
- Extensive experiments demonstrate that our method outperforms the SOTA methods in VG tasks, with only **2.04%** tunable parameters. Moreover, SwimVG offers significant computing efficiency advantages.

II. RELATED WORK

A. Visual Grounding

Visual grounding (VG) [9], [13], [14], [19], [24]–[26] aims to identify and localize regions within images that correspond to given text descriptions. There are many extensions of VG in other fields, such as Remote Sensing VG [27]–[29]. Early visual grounding methods, given their resemblance to detection tasks, initially aligned with the prevailing object detection architectures. These architectures evolve from the initial two-stage designs to the recently one-stage methods. Two-stage designs methods [30]–[32] follow a two-stage pipeline that 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. With the introduction of ViT, the Transformer-based methods [2], [9], [14], [33]–[37] further propose an end-to-end framework which reformulate the prediction process as a regression problem. Most recently, grounding multimodal large language models [38]–[40] have propelled the state-of-the-art (SOTA) performance, these works require a large amounts of in-domain and other domain datasets. Despite the transformer-based models exhibiting ideal performance in VG, most methods involve fully fine-tuning the text and visual branches separately, followed by a heavyweight vision-language encoder for simple multimodal fusion. This not only makes it difficult to focus on the areas most relevant to the text description but is also inefficient.

B. Parameter-Efficient Transfer Learning

Transfer learning aims to adapt pre-trained models to specific tasks or datasets. With the growth of model sizes and the

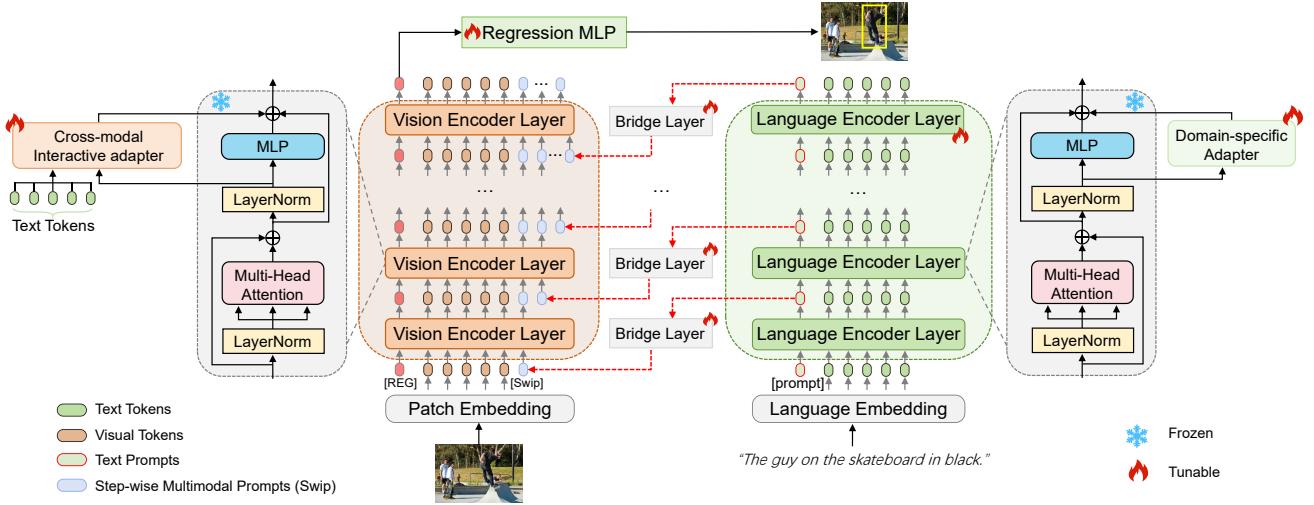


Fig. 2. Overall architecture of the proposed SwinVG, which freezes the pre-trained vision encoder and language encoder. SwinVG integrates step-wise multimodal prompts (Swip) and cross-modal interactive adapters, which bridges the visual and language encoders, ensuring the visual encoder concentrates on the text-relevant areas.

complexity of the specific tasks, fully fine-tuning paradigm demands significant computational resources. To address these challenges, researchers in the NLP and CV domains have explored PETL methods [18], [41]–[44]. One method, known as Prompt Tuning [15], [16], [45], involves the introduction of trainable tokens at the input space, thereby learning task-specific representations. Adapter-like methods [43], [46] involve inserting additional trainable weights, such as Multi-Layer Perceptrons (MLPs) equipped with activation functions and residual connections, within the network architecture to enhance transfer learning capabilities. Meanwhile, LoRA-like methods [42] adjust pre-trained models by using the idea of low-rank matrix decomposition, and only trains the parameters of the low-rank matrix. LoRA-like methods have been proposed in the field of natural language processing for Large Language Models (LLM) such as GPT-4 [47], LLaMA2 [48], and GLM-4 [49]. By focusing on updating only a small subset of parameters, PETL methods effectively simulate the fine-tuning of the entire model’s parameters without directly modifying them. Recently, some pioneering works like MaPPER [50], HiVG [13], DARA [14] and M²IST [51] sought to utilize adapters to adapt pre-trained models to visual grounding. However, they all use a burdensome vision-language module for multimodal fusion, which is not an efficient enough method.

III. METHOD

Our method is designed to enhance the generalization capabilities of pre-trained models in the realm of visual grounding efficiently. This is achieved through step-wise multimodal prompts, light domain-specific adapters, and cross-modal interactive adapters. Fig. 2 shows the overall architecture of our proposed SwinVG framework.

A. Text & Image Backbone

Given an image and a text, we extract their features through the image encoder and text encoder, respectively.

Text Encoder. Given the input text expression T with a length of L , we utilize the pre-trained text branch of CLIP [52] for extracting text features. The text expression is firstly converted into a one-hot vector. Subsequently, each one-hot vector is tokenized into a series of linguistic tokens, 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. The input embeddings $\hat{T} \in \mathbb{R}^{L \times C_t}$, where $\hat{T} = [t^1, t^2, \dots, t^L]$, and C_t is the dimension of text embeddings.

Visual Encoder. We use DINOv2 [53] as the visual backbone. The model involves training the Vision Transformer (ViT) model [54] on the extensive LVD-142M dataset, and employs a self-supervised learning strategy. This method allows the model to extract powerful visual features, thereby offering remarkable performance in various downstream tasks. Given an input image $I_0 \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the image is initially divided into N non-overlapping patches, which are then linearly projected into D -dim patch embeddings $I'_0 \in \mathbb{R}^{D \times C_v}$. Motivated by TransVG [9] and HiVG [13] appending a learnable [REG] in vision-language transformer, we also adopt a learnable [REG] token to directly predict the 4-dim coordinates of a referred object. Unlike the previous method, we omit the complex vision-language fusion structure and directly pre-append the [REG] token to I'_0 , and the token is processed by the visual encoder layer gradually.

As the vision and language backbones contain most model parameters and have acquired rich knowledge during pre-training, we attempt to freeze them during fine-tuning. This strategy allows for a more efficient allocation of computational resources and focuses the learning on the adjustments of other modules.

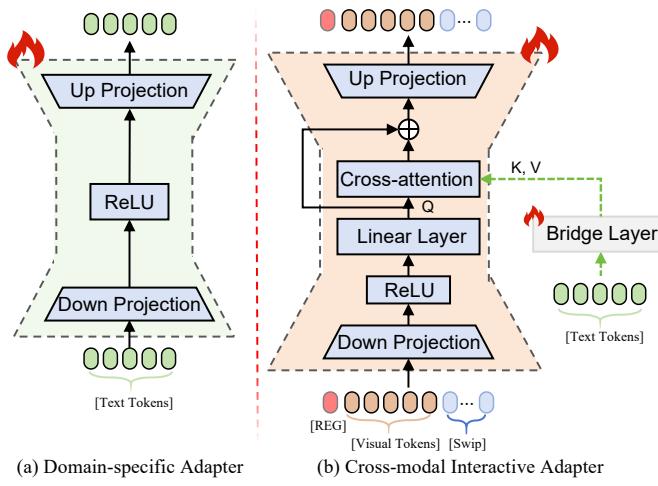


Fig. 3. The Domian-specific adapter and cross-modal interactive adapter.

B. Step-wise Multimodal Prompting

The intuitive idea of achieving token-level multimodal alignment is to directly concatenate text tokens and vision tokens together for learning. However, an increase in the input length will bring about a computational burden. To efficiently establish token-level multimodal alignment, we design step-wise multimodal prompts, and introduce these learnable tokens in the layers of both vision and language branches from shallow to deep layer. This means that these tokens are added to transformer layers in a hierarchical way. The hierarchical multimodal prompts utilize the knowledge embedded in pre-trained models to effectively learn task-relevant cross-modal representations.

Text Prompting. To learn the represent the global textual representation, a learnable token $p \in \mathbb{R}^H$ is introduced in the text encoder. The input embeddings \hat{T} is converted to \hat{T}' , follow the form $[p, t^1, t^2, \dots, t^L]$. The new token is further processed by each transformer block of the language encoder \mathcal{L}_i . This process can be formulated as below:

$$[p_i, T_i] = \mathcal{L}_i([p_{i-1}, T_{i-1}]) \quad i = 1, 2, \dots, l. \quad (1)$$

where p_i and T_i represent the prompt and text embeddings processed by the i -th language encoder layer, respectively. The l refers to the depth of the language encoder. The prompt is initialized by Xavier initialization.

Step-wise Multimodal Fusion. To efficiently fuse textual and visual semantics step-by-step, we gradually convey the prompt processed by the text encoder to vision encoder. Due to the different feature dimensions of the text and vision encoder, we need to adjust the features to the same dimension. Therefore, we design a bridge layer to transport text features, making them adaptable for visual branch. For the visual encoder layer \mathcal{V}_{i-1} , we introduce the token from the \mathcal{L}_i language encoder layer to the layer \mathcal{V}_{i-1} of vision encoder. Since the prompt added to the visual token set is initialized by global textual semantics, when the prompt is introduced into the corresponding visual layer, it can facilitate multimodal fusion step by step, namely step-wise multimodal prompt (Swip).

Each Swip is further processed by the deeper visual layer. The process can be formalized as:

$$m_i = p_i \mathbf{W}_{bridge} \quad (2)$$

$$[m_i^0, \dots, m_i^{i-1}, \mathbf{V}_i] = \mathcal{V}_i([m_{i-1}^0, \dots, m_{i-1}^{i-1}, \mathbf{V}_{i-1}]) \quad (3)$$

where $\mathbf{W}_{bridge} \in \mathbb{R}^{C_t \times C_v}$ is the weights of bridge layer. The m_i are multimodal prompts transformed from text prompts. The n refers to the depth of vision encoder, and m_i^0 represents the 0-th swip token processed by the i -th vision encoder layer. The \mathbf{V}_i is vision emdeddings processed by the i -th vision encoder layer.

C. Cross-modal Interactive & Domain Adaption

To efficiently transfer pre-trained text semantics knowledge to visual grounding, and further facilitate multimodal interaction, we introduce domain-specific adapters in the text encoder and cross-modal interactive adapters in vision encoder, respectively.

Cross-modal Interactive Adapter. As shown in Fig. 3(b), we design a Cross-modal Interactive Adapter (CIA) to make the interaction of modal information between the visual encoder and text encoder, which enhances the capability of multimodal fusion while fixing the backbone parameters. The main difference between the design of CIA and previous adapters lies in the integration of a cross-modal attention module. To ensure the lightweight and efficiency of whole structure, we firstly adopt a down-projection to transform the visual features to low-rank features. CIA module is inserted between the activation and up-projection layers. Similar to step-wise multimodal fusion, text features should be converted by bridge layer to dimensions that match the visual branches. Given vision features f_i^v process by the Multi-Head Attention (MHA) of the layer \mathcal{V}_i , and text features f_i^t process by the MHA of the layer \mathcal{L}_i , this process can be formulated as below:

$$c_i^t = f_i^t \mathbf{W}_{bridge} \quad (4)$$

$$\begin{aligned} f_{down}^v &= f_i^v \mathbf{W}_{down}, \\ f_{act}^v &= \text{ReLU}(f_{down}^v), \\ f_l^v &= f_{act}^v \mathbf{W}_{linear}. \end{aligned} \quad (5)$$

$$\text{MHCA}(f_l^v, c_i^t) = \text{Softmax}\left(\frac{f_l^v W_q c_i^t W_k}{\sqrt{C}}\right)(c_i^t W_v) \quad (6)$$

$$f_{up} = (f_l^v + \text{MHCA}(f_l^v, c_i^t)) \mathbf{W}_{up} \quad (7)$$

$$\text{CIA}(f_i^v, c_i^t) = f_i^v + s_{vt} \cdot f_{up}. \quad (8)$$

Here, $\mathbf{W}_{down} \in \mathbb{R}^{C_v \times C_d}$ and $\mathbf{W}_{up} \in \mathbb{R}^{C_d \times C_v}$ are the weights of down- and up-projection layers, and s_{vt} is the scaling factor for multimodal fusion. The MHCA is Multi-Head Cross-Attention module in CIA.

Domain-specific Adapter. Due to fully freezing of the text backbone, there exists a gap between pre-trained model and

TABLE I

COMPARISON WITH LATEST SOTA METHODS ON REFCOCO/+G FOR VISUAL GROUNDING. "RN50", "RN101", "DN53", AND "SWIN-S" REPRESENT RESNET-50 [55], RESNET-101 [55], DARKNET-53 [56], AND SWIN-TRANSFORMER SMALL, RESPECTIVELY. \dagger 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 WHOLE MODEL. THE BOLDFACE DENOTES THE BEST PERFORMANCE WHILE THE UNDERLINE INDICATES THE SECOND BEST.

Methods	Venue	Backbone	Tuned/Total param.	RefCOCO			RefCOCO+			RefCOCOg			Flickr30K
				val	testA	testB	val	testA	testB	val-g	val-u	test-u	test
Full Fine-tuning													
MAttNet [24]	CVPR'18	RN101/LSTM	100%	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27	-
RvG-Tree [31]	TPAMI'19	RN101/LSTM	100%	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51	-
NMTTree [30]	ICCV'19	RN101/LSTM	100%	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44	-
FAOA [25]	ICCV'19	DNS3/LSTM	100%	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.26	68.71
ReSC-Large [57]	ECCV'20	ND53/BERT-B	100%	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20	69.28
TransVG [9]	ICCV'21	RN50+DETR/BERT-B	100%	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44	78.47
QRNet [12]	CVPR'22	Swin-S/BERT-B	100%	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03	72.52	81.95
Dynamic-MDETR \dagger [11]	TPAMI'23	CLIP-B	100%	85.97	<u>88.82</u>	80.12	74.83	81.70	63.44	72.21	74.14	74.49	81.89
PFOS [58]	TMM'22	DNS3/BERT-B	100%	77.37	80.43	72.87	63.74	68.54	55.84	61.46	67.08	66.35	-
SeqTR [35]	ECCV'22	DNS5/BiGRU3	100%	81.23	85.00	76.08	68.82	75.37	58.78	-	71.35	71.58	81.23
Word2Pix [59]	TNNLS'22	RN101+DETR/BERT-B	100%	81.20	84.39	78.12	69.46	76.81	61.57	-	70.81	71.34	-
YORO \dagger [60]	ECCV'22	ViLT	100%	82.90	85.60	77.40	73.50	78.60	64.90	-	73.40	74.30	-
TransVG++ [10]	TPAMI'23	ViT-Det/BERT-B	100%	86.28	88.37	80.97	75.39	80.45	66.28	73.86	76.18	76.30	81.49
CLIP-VG [19]	TMM'23	CLIP-B	100%	84.29	87.76	78.43	69.55	77.33	57.62	72.64	73.18	72.54	81.99
JMRI [37]	TIM'23	CLIP-B	100%	82.97	87.30	74.62	71.17	79.82	57.01	69.32	71.96	72.04	79.90
PTP2R-BLIP [61]	TPAMI'23	BLIP	100%	81.83	86.44	74.30	76.65	<u>82.14</u>	<u>67.38</u>	-	-	-	-
VG-LAW [2]	CVPR'23	ViT-Det/BERT-B	100%	86.06	88.56	<u>82.87</u>	75.74	80.32	66.69	-	75.31	75.95	-
MGCross [62]	TIP'24	RN101/BERT-B	100%	85.10	88.23	80.08	74.44	79.48	65.21	74.50	<u>77.25</u>	75.78	75.18
TransCP [63]	TPAMI'24	RN50/BERT-B	100%	84.25	87.38	79.78	73.07	78.05	63.35	72.60	-	-	80.04
LGR-NET [26]	TCSV'T24	Swin-S/BERT-B	100%	85.63	88.24	82.69	75.32	80.60	68.30	<u>75.48</u>	76.82	<u>77.03</u>	81.97
ScanFormer [64]	CVPR'24	ViLT	100%	83.40	85.86	78.81	72.96	77.57	62.50	74.10	-	74.14	68.85
Parameter-efficient Transfer Learning													
DARA [14]	ICME'24	RN50+DETR/BERT-B	7.14%	81.16	82.76	76.72	65.58	69.83	57.22	67.21	69.22	67.67	-
MaPPER [50]	EMNLP'24	DINOv2/BERT-B	6.2%	86.03	88.90	81.19	74.92	81.12	65.68	74.60	76.32	75.81	-
HiVG [13]	MM'24	CLIP-B	23.04%	87.32	89.86	83.27	78.06	84.81	68.11	-	78.29	78.79	82.11
SwimVG (Ours)	-	DINOv2/CLIP-B	2.04%	88.29	90.37	84.89	77.92	83.22	69.95	79.10	80.14	79.69	83.10

visual grounding. To address the issue, we incorporate domain-specific adapters (DoSA) to improve the text encoder for domain understanding. As shown in Fig. 3(a). Compared to the CIA adapter, the domain-specific adapter adopts a more straightforward design, focusing on learning text representation efficiently without complex structure. This neat yet effective approach ensures efficient processing of text semantics while maintaining compatibility with the overall model architecture. By taking advantage of these enhanced features, the model facilitates aligning visual and linguistic features. Specifically, the domain-specific adapter follows a standard "Down-ReLU-Up" structure to bridge the gap between pre-trained knowledge and visual grounding. Given the text features f_i^t processed by the Multi-Head Attention (MHA) of the layer \mathcal{L}_i , the learning process can be formalized as:

$$\begin{aligned} t_{down} &= f_i^t \mathbf{W}_{down}, \\ t_{act} &= \text{ReLU}(t_{down}), \\ t_{up} &= t_{act} \mathbf{W}_{up}, \end{aligned} \quad (9)$$

$$\text{DoSA}(f_i^t) = f_i^t + s_t \cdot t_{up}, \quad (10)$$

where $\mathbf{W}_{down} \in \mathbb{R}^{C_t \times C_d}$ and $\mathbf{W}_{up} \in \mathbb{R}^{C_d \times C_t}$ are the weights of down- and up-projection layers, and s_t is the scaling factor of domain-specific adapters. In this way, DoSA can refine the rich pre-trained language representations into more fine-grained representations for the VG domain during fine-tuning.

D. Prediction Head

Followed by HiVG [13] and TransVG++ [10], a regression block with a MLP and a linear layer are adopted to perform box coordinates prediction. Given the [REG] token from the last layer of vision encoder, the regression block generates the 4-dim bounding box coordinates.

E. Training Objectives

Following the previous work [9], [14], the L1 loss and Generalized IoU (GIoU) loss are used between the predicted bounding box coordinates $\tilde{b} = (\tilde{x}, \tilde{y}, \tilde{w}, \tilde{h})$ and the ground truth $b = (x, y, w, h)$, the training objective for VG is defined as follows:

$$\mathcal{L}_{\text{rec}} = \lambda_1 \mathcal{L}_{\text{L1}}(b, \tilde{b}) + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}(b, \tilde{b}), \quad (11)$$

where $\mathcal{L}_{\text{L1}}(\cdot)$ and $\mathcal{L}_{\text{giou}}(\cdot)$ represent L1 loss and GIoU loss [65], respectively. The λ_1 and λ_{giou} are the weight coefficient to balance the two detection loss functions.

IV. EXPERIMENTS

In this section, we will give a detailed experimental analysis of the whole framework, including the datasets, evaluation protocol, implementation details, comparisons with the state-of-the-art methods, and ablation analysis.

A. Experimental Setup

Datasets. To verify the effectiveness and efficiency of our method, we have conducted comprehensive experiments on

the RefCOCO [66], RefCOCO+ [66], RefCOCOg [21], [22] and Flickr30K Entities [23] datasets, all of which are widely used as benchmarks for visual grounding.

- **RefCOCO** features 19,994 images with 50,000 referred objects and 142,210 expressions. The dataset is divided into four subsets, consisting of 120,624 train, 10,834 validation, 5,657 test A, and 5,095 test B samples, respectively. The average length of the expressions is 3.6 words, and each image contains a minimum of two objects.
- **RefCOCO+** with similar content but richer expressions, includes 19,992 images with 49,856 referred objects and 141,564 referring expressions. The dataset is divided into four subsets: 120,624 train, 10,758 validation, 5,726 test A, and 4,889 test B samples. Notably, the RefCOCO+ dataset has been constructed to be more challenging than the RefCOCO dataset by excluding certain types of absolute-location words. The average length of the expressions is 3.5 words, including the attribute and location of referents.
- **RefCOCOg**, unique for its detailed annotations and longer referential expressions, contains 25,799 images with 49,856 objects. There are two commonly used split protocols for this dataset. One is RefCOCOg-google [21], and the other is RefCOCOg-umd [22]. We report our performance on both RefCOCOg-google (val-g) and RefCOCOg-umd (val-u and test-u) to make comprehensive comparisons. The average length of expressions within the dataset is 8.4 words, including both the attributes and the locations of the referents. This rich detail description facilitates a more nuanced understanding of the visual grounding tasks, as it captures the intricacies of how objects are referenced in various contexts.
- **Flickr30K Entities** [23], is an enhanced version of the original Flickr30K [67], fortified with the addition of short region phrase correspondence annotations. This expansion yields a collection of 31,783 images, encompassing 427,000 referred entities. Following the previous studies [13], [40], we have divided the dataset into 29,783 images for training, 1,000 for validation, and another 1,000 for testing purposes.

Evaluation Metrics. We follow the previous research that employs top-1 accuracy (%) as the evaluation metric for visual grounding. Specifically, a prediction is deemed accurate only when its Intersection-over-Union (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-L/14 [53], while the language encoder uses CLIP-B [52]. The resolution of the input image is 224x224. The DINOv2-L/14 model processes tokens with a feature dimension of 768, while the CLIP-B model handles tokens with a feature dimension of 512. All prompts use Xavier initialization, and all adapters are initialized with Kaiming normal initialization. The bottleneck dimension C_d for both CIA and domain-specific adapters is 56, and more dimension

comparisons can be seen in Table VII. The batchsize for training is 32. For fair comparisons, other PETL methods in Tab. II use the same base architecture and original hyperparameters, and keeping the vision and language encoder frozen. For RefCOCO [20], RefCOCOg [21], [22], and Flickr30K Entities [23] datasets, the entire network is trained for 65 epochs using the AdamW optimizer. While for RefCOCO+ [20] dataset, the network is trained for 90 epochs. Note that most mainstream methods train RefCOCO/RefCOCOg/Flickr30K Entities for 90 epochs and RefCOCO+ for 180 epochs, which demonstrates the higher efficiency of our SwimVG. We conduct all experiments on one A800 GPU.

B. Main Results

We compare our SwimVG comprehensively with a series of previous visual grounding (VG) methods. The main experimental results are displayed in Tab. I. We can notice from these results that SwimVG reaches the best accuracy and also ensures parameter efficiency compared with all other methods, which validates its effectiveness and efficiency.

Effectiveness. As Tab. I shown, on the three commonly challenging benchmarks, SwimVG outperforms all traditional full fine-tuning methods. Compared to DARA [14], a parameter-efficient transfer learning method, we achieve an average accuracy improvement of 10.85% on the three benchmarks. Notably, even compared to some methods that are pre-trained on the RefCOCO+/g and Flickr30K Entities (indicated by † in Tab. I), our SwimVG model achieves the highest scores across all evaluation tasks, with particularly strong performance on the RefCOCO+, which present greater challenges compared to RefCOCO.

Efficiency. Tab. I clearly illustrates that SwimVG not only achieves the best performance, but also highlights its huge advantages in parameter efficiency. SwimVG reduced the tunable backbone parameters by 97.96% compared to the traditional full fine tuning method. In order to verify more efficient aspects such as training and inference time, experimental results on the mainstream methods using the conventional VL transformer, and the other PETL methods are shown in Tab. III. It can be seen that SwimVG achieves significant energy efficiency advantages.

C. Comparison with Other PETL Methods

Details of Baseline PETL Methods.

This section furnishes additional details of the PETL baselines employed in our primary manuscript. Notably, all these baselines follow the same base architecture.

- **AdaptFormer [18]:** We add adapters in parallel to MHA and FFN in both Vision Encoder and Language Encoder. Following the original work, we set the same bottleneck dimensions of AdaptFormer for both vision and language branch.
- **LoRA [42]:** We incorporate trainable matrices in parallel to the weight matrices in MHA and FFN in both Vision Encoder and Language Encoder. We have set the same bottleneck dimensions for both the vision and language branches of LoRA, following the original setup.

TABLE II
COMPARISON WITH PETL METHODS USING THE SAME BACKBONE AS SWIMVG ON REFCOCO, REFCOCO+ AND REFCOCOG. “PARAM.” INDICATES THE NUMBER OF TUNABLE PARAMETERS IN THE PRE-TRAINED ENCODERS.

Methods	Venue	RefCOCO			RefCOCO+			RefCOCOG		
		val	testA	testB	val	testA	testB	val-g	val-u	test-u
AdaptFormer [18]	NeurIPS’22	81.75	83.14	76.73	72.05	76.61	64.26	70.19	70.93	72.36
LoRA [42]	ICLR’22	82.43	84.51	77.32	72.66	77.13	64.85	71.27	72.16	73.23
UniAdapter [68]	ICLR’24	85.76	88.31	81.84	74.95	78.75	65.97	73.68	74.72	74.98
DAPT [44]	CVPR’24	85.33	87.52	81.06	74.33	78.66	65.54	74.02	75.26	75.47
SwimVG	-	88.29	90.37	84.89	77.92	83.22	69.95	79.10	80.14	79.69

TABLE III

EFFICIENCY COMPARISON. THE RESULTS ARE OBTAINED ON REFCOCO DATASET. “–” INDICATES THAT THE MODEL’S CODE IS NOT PUBLICLY AVAILABLE, AND THEIR RESULTS ARE NOT AVAILABLE.

Model	update/all param.	update ratio	train time (epoch/min)	testA time (s)	testA Acc.↑
Full Fine-tuning					
TransVG	159.4/159.4M	100%	52	95	82.67
QRNet	281.4/281.4M	100%	62	111	85.85
VG-LAW	158.7/158.7M	100%	–	–	88.56
TransVG++	171.13/171.13M	100%	–	–	88.37
PETL Methods					
DARA	11.61/162.61M	7.14%	33	96	82.76
LoRA	17.15/392.15M	4.37%	61	127	84.51
AdaptFormer	14.85/389.85M	3.81%	57	125	83.14
Uniadapter	29.16/404.16M	7.21%	65	131	88.31
DAPT	26.69/401.69M	6.64%	64	129	87.52
HiVG	49.40/214.40M	23.04%	–	–	89.86
SwimVG(ours)	7.65/375.13M	2.04%	40	65	90.37

- UniAdapter [68]:** We add UniAdapter in both Vision Encoder and Language Encoder, according to their basic designs.
- DAPT [44]:** We insert Dynamic Adapters in parallel to the weight matrices in MHA and FFN in both Vision Encoder and Language Encoder, and use their task-agnostic feature transform strategy. Other sets such as bottleneck dimensions are same as the DAPT.

We conduct experiments comparing our SwimVG with other parameter-efficient transfer learning (PETL) methods. To ensure fairness, we retain the original parameter settings from previous methods. As these PETL methods lack the capability of multimodal fusion, we complement them with the traditional VL transformer for cross-modal understanding, thereby enabling a direct comparison with our SwimVG. Tab. II illustrates that SwimVG outperforms other PETL methods on all three benchmarks. Through introducing step-wise multimodal prompts and cross-modal interactive adapters, SwimVG enhances the modeling of the vision-text alignment capability. Previous PETL methods lack this ability, rendering them less effective for VG tasks. This also proves that the multimodal fusion mechanism in SwimVG is more efficient than the VL transformer. To summarize, by the specific design for the VG domain, SwimVG achieves superior performance with only **2.04 %** tunable parameters.

D. Convergence Analysis

Figure 5 shows a comparison of the convergence epoch between SwimVG and other models. It is observed that DARA and TransVG converge around epoch 85, while CLIP-VG converges at approximately epoch 105. In contrast, SwimVG achieves convergence at around epoch 65. This demonstrates the efficiency of our method, as fewer training epochs are required, thereby reducing training costs. In addition, we have also visualized the convergence comparison of SwimVG across the RefCOCO, RefCOCOG-u, RefCOCOG-g, and Flicker 30K datasets. Figure 6 indicates that convergence is achieved around epoch 65 for all these datasets.

E. Ablation Study

Effectiveness of Multimodal Interaction in SwimVG. We assess the impact of step-wise multimodal prompts (Swip) and cross-modal interactive adapters (CIA) by performing an ablation study, and report the results on RefCOCOG-u validation and test datasets. Considering the substantial number of parameters occupied by the encoder, we freeze all the encoder parameters during fine-tuning for efficiency. From Tab. IV, it is evident that only introducing the Swip yields a ideal results (Tab. IV (a)). Only by using the CIA for cross-modal fusion can achieve better results (Tab. IV (b)). Compared with the previous methods using the traditional vision-language encoder, such as TransVG [9], DARA [14] in Tab. I, it shows that we can achieve the better results using only Swip or CIA. Tab. IV (c) indicates that incorporating Swip and CIA for multimodal fusion results in an average improvement of 3.49% across the RefCOCOG-u, achieving the best performance among these ablation variants. Swip achieves progressive multimodal fusion by gradually introducing linguistic information, while CIA explores deeper correlations by enhancing cross-modal interaction. Combining the two can simultaneously promote multimodal fusion in terms of breadth and depth.

Effectiveness of Domain-specific Adapters. Because the text encoder is pre-trained on a general domain, freezing the entire text backbone restricts the specific language understanding in visual grounding domain, thereby weakening the proper interaction between text and vision semantics. To enable the domain text semantics to interact with the visual encoder efficiently, we adopt domain-specific adapters to learn the domain knowledge, thus making the text encoder match with visual grounding. Tab. V shows that domain-specific adapters efficiently transfer the language knowledge of the pre-trained

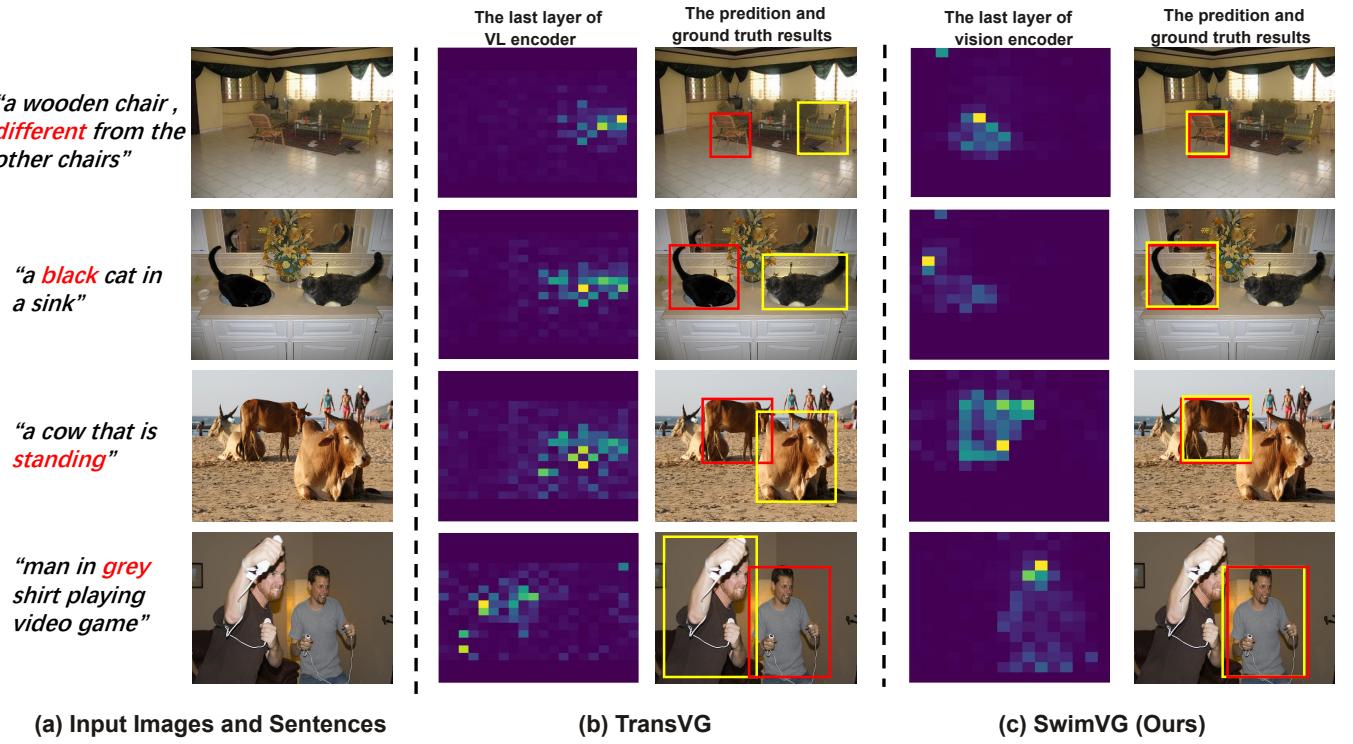


Fig. 4. Visualizations of attention maps, prediction results (yellow bounding boxes) and ground truth (red bounding boxes).

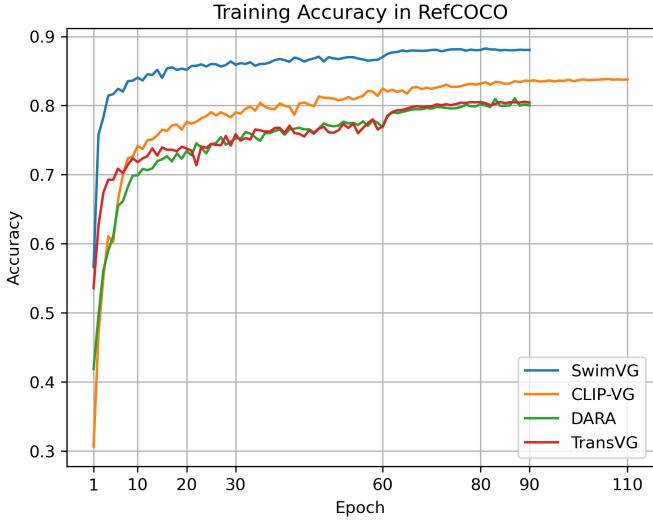


Fig. 5. The convergence comparison between SwimVG and other SOTA models on RefCOCO.

model to VG domain, further improving an average improvement of 4.39% across the RefCOCOG-u.

Effects of Different Insertion Positions of SwimVG. To determine the optimal configuration of the Cross-modal Interactive Adapter (CIA) and Text Adapter, we conducted an ablation study varying both different layers and the dimensions of the adapters. Firstly, we evaluated the impact of different adapter layers. In this experiment, the visual CIA and the Text Adapter were inserted at the same layers. From Table VI, we can see that: (1) Only inserting three layers for vision and text encoder can bring great performance (Table VI (a));

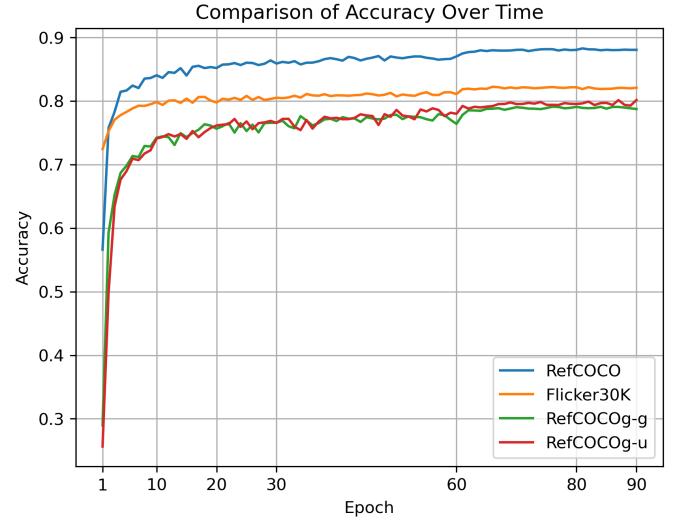


Fig. 6. The convergence comparison of SwimVG on RefCOCO, RefCOCOG and Flicker 30K datasets.

(2) observing Table VI (b), (c), and (d), it can be seen that inserting CIA later in the vision encoder can exhibit better performance; (3) from the observation of Table VI (e) and (f), it is evident that inserting text adapter later in the text encoder results in a minor performance decline; (4) adding adapters from 13 layers to 24 layers not only reduces performance but also increases the tunable parameters. This might be because the visual backbone is more likely to adapt to the VG domain at deeper layers, while the text needs to adapt from the shallow layers to the deep layers. It should be noted that the text encoder is composed of 12 layers, while the vision encoder

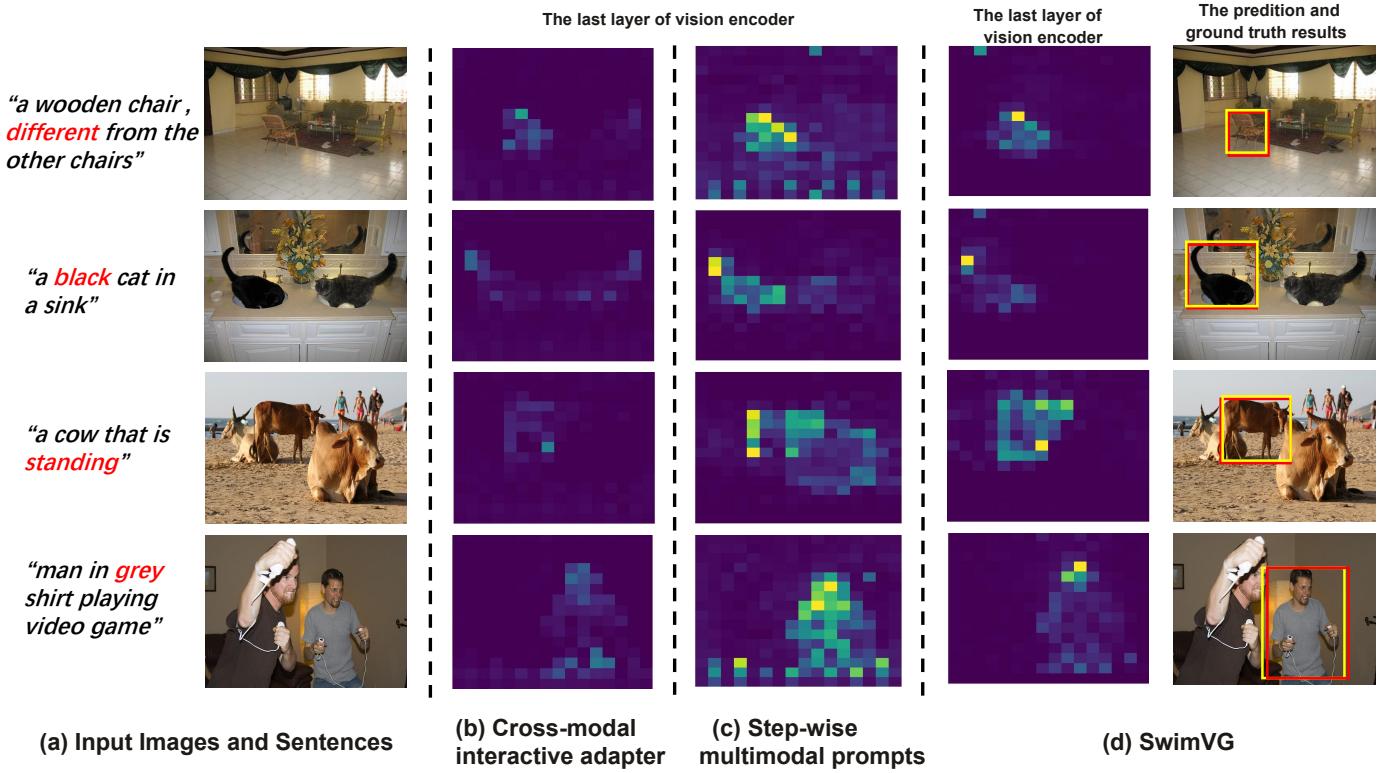


Fig. 7. The visualizations of attention maps from vision encoder with different strategies of SwimVG. Red bounding boxes represent ground truth, and yellow bounding boxes are prediction results.

TABLE IV

ABLATIONS OF MULTIMODAL INTERACTION IN SWIMVG ON REFCOCOG-U [66] DATASET. NOTE THAT THE VISUAL AND TEXT ENCODER ARE FROZEN IN THE ABLATION STUDIES.

	Step-wise Multi. Prompts	Cross-modal Inter. Adapters	Updated Params.	RefCOCOG val-u	RefCOCOG test-u
(a)	✓		6.30M	71.32	70.06
(b)		✓	1.00M	72.22	71.86
(c)	✓	✓	7.30M	75.57	75.48

TABLE V

EFFECTIVENESS OF DOMAIN-SPECIFIC ADAPTERS. (A) REPRESENTS INTRODUCING SWIP AND CIA IN SWIMVG.

#	Domain-specific Adapters	Updated Params.	RefCOCOG val-u	RefCOCOG test-u
(a)		7.30M	75.57	75.48
(b)	✓	7.65M	80.14	79.69

comprises 24 layers.

Effects of Different Hyper-parameter Settings of SwimVG.

We first ablate the bottleneck dimensions C_d of all adapters (see Table VII (a,b,c)), and follow the design shown in Table VII (a). C_d determines the number of tunable parameters introduced by SwimVG. As shown in Table VII, higher C_d introduces more parameters, and the performance consistently increases when C_d increases up to 56. C_d 128 exhibits

TABLE VI

ABLATION STUDY OF DIFFERENT CONFIGURATIONS OF CROSS-MODAL INTERACTIVE ADAPTERS AND TEXT ADAPTERS. FOR THE “POSITION”, WE LIST THE I-TH LAYERS THAT INSERT ADAPTERS IN THE BACKBONE.

#	Position		Params	RefCOCOG	
	text	vision		val-u	test-u
(a)	4,8,12	8,16,24	6.67M	75.26	74.78
(b)	2,4,6,8,10,12	4,8,12,16,20,24	7.65M	78.65	72.54
(c)	2,4,6,8,10,12	14,16,18,20,22,24	7.65 M	79.28	78.62
(d)	2,4,6,8,10,12	19,20,21,22,23,24	7.65M	80.14	79.69
(e)	7,8,9,10,11,12	19,20,21,22,23,24	7.65M	78.90	78.06
(f)	7,8,9,10,11,12	14,16,18,20,22,24	7.65M	79.06	78.43
(g)	2,4,6,8,10,12	13-24	8.65M	79.51	78.39

TABLE VII
EFFECTIVENESS OF DIFFERENT BOTTLENECK FOR ALL ADAPTERS.

#	Bottleneck dimensions	Params. (M)	RefCOCOG val-u	RefCOCOG test-u
(a)	32	7.05	78.65	78.13
(b)	40	7.24	79.67	78.78
(c)	56	7.65	80.14	79.69
(d)	64	7.87	79.12	78.52
(e)	128	9.76	80.18	79.43

TABLE VIII
COMPARISON OF THE CONTRIBUTION LEVELS OF DIFFERENT BACKBONES.

Mehthods	Vision Backbone	Language Backbone	RefCOCO		
			val	testA	testB
TransVG [9]	RN101+DETR	BERT-Base	81.02	82.72	78.35
TransVG [9]	DINOv2-L	BERT-Base	85.11	87.36	80.97
TransVG [9]	DINOv2-L	CLIP-Base	85.55	86.79	80.28
SwimVG	DINOv2-L	CLIP-Base	88.29	90.37	84.89

TABLE IX
COMPARISON OF THE MORE EVALUATION METRICS.

Mehthods	Pr@0.6(RefCOCO)			Pr@0.8(RefCOCO)		
	val	testA	testB	val	testA	testB
MaPPER [50]	82.23	86.03	76.11	66.62	72.63	57.50
SwimVG	85.26	87.33	80.61	68.86	72.83	63.04

considerable performance, but its tunable parameter count is about twice that of C_d 56. Thus, we select the C_d as 56. This indicates that a small bottleneck may not provide sufficient adaptation capabilities, while a large dimension may lead to over-adaptation. An intermediate dimension can achieve a better adaptation to the VG domain.

The contribution degree of different pre-trained models. To facilitate the analysis of the contribution of different backbones to performance, we excluded the SwimVG method and compared different backbones based on TransVG [9]. We selected ResNet101+DETR and DINOv2-L as the vision backbone and chose the mainstream BERT-Base and the text encoder in CLIP-Base as the text backbone. As seen in Table VIII, the vision backbone has a relatively large impact on visual grounding, whereas the text backbones have a relatively small impact. Under the same backbone, our method outperforms TransVG, which indicates that our multimodal fusion strategy is highly effective.

F. More Evaluation Metrics

We compared more challenging evaluation metrics, such as the prediction accuracy when $\text{IoU} > 0.6$ (Pr@0.6) and Pr@0.8. Under the same metrics, we compared the latest MaPPER [50]. As seen in Table IX, SwimVG outperforms the latest MaPPER under both the settings of Pr@0.6 and Pr@0.8.

G. Qualitative Results

The comparison of multimodal fusion strategy. To verify that the multimodal fusion strategy of SwimVG is superior to the traditional vision-language transformer (VL encoder), we visualize the attention maps from the last layer of vision encoder in SwimVG. Due to the suboptimal multimodal fusion methods employed by other mainstream approaches, namely the visual language transformer (VL encoder), which lack open-source code or checkpoints, we opt to visualize the last layer of the VL encoder from TransVG. As shown in Fig.4, TransVG fails to pay sufficient attention to text-relevant regions in images. For example, TransVG lacks the alignment ability of “*different*”, “*black*”, and “*standing*” with images. The comparison with TransVG demonstrates the ability of our proposed SwimVG to focus more on the text-relevant regions,

and our multimodal fusion strategy is superior to the traditional VL encoder.

The effectiveness of CIA and Swip. In this section, we present more visualization of the attention maps from the vision encoder under different mixing strategies. As depicted in Figure 7, we can see that: (1) introducing either cross-modal interactive adapters (CIA) or step-wise multimodal prompts (Swip) facilitates the interaction between the vision and language encoders. (Figure 7 (b,c)); (2) compared to CIA, the attention map of only introducing is slightly scattered (Figure 7 (b,c)); integrating CIA and Swip can further enhances the facilitation of cross-modal interaction (Figure 7 (d)). The interaction between the vision and language encoder, facilitated by CIA and Swip, allows the model to focus more effectively on the referred objects in diverse expression cases.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper, we aims at improving both the effectiveness and efficiency of visual-text alignment. We propose SwimVG by the foundational design of step-wise multimodal prompts (Swip) and cross-modal interactive adapters (CIA). SwimVG integrates a novel multimodal fusion strategy of token-level Swip and weight-level CIA to enable the visual encoder can concentrate on the text-relevant regions. Extensive experiments and ablation studies have validated the high effectiveness of our method. Our proposed framework significantly outperforms the baseline and achieves comparable results with the state-of-the-art methods while tiny parameter budget.

B. Future Work

In the future, implementing our SwimVG in real-world applications is a challenging and meaningful direction. Currently, our SwimVG has only been evaluated on benchmark datasets. However, its performance against datasets from different domains remains unknown. In addition, the efficient multi-modal fusion strategies of SwimVG can be verified on other multimodal tasks, such as visual question answering and video caption. Motivated by efficient Multimodal Large Language Model [69], we will explore efficient training and inference model for visual grounding.

REFERENCES

- [1] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “Mdetr-modulated detection for end-to-end multi-modal understanding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.
- [2] W. Su, P. Miao, H. Dou, G. Wang, L. Qiao, Z. Li, and X. Li, “Language adaptive weight generation for multi-task visual grounding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10857–10866.
- [3] Y. Qiao, C. Deng, and Q. Wu, “Referring expression comprehension: A survey of methods and datasets,” *IEEE Transactions on Multimedia*, vol. 23, pp. 4426–4440, 2020.
- [4] L. Xiao, X. Yang, X. Lan, Y. Wang, and C. Xu, “Towards visual grounding: A survey,” *arXiv preprint arXiv:2412.20206*, 2024.
- [5] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, “Embodied question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1–10.

- [6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [7] X. Zhang, L. Wang, G. Zhang, T. Lan, H. Zhang, L. Zhao, J. Li, L. Zhu, and H. Liu, "Ri-fusion: 3d object detection using enhanced point features with range-image fusion for autonomous driving," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2022.
- [8] A. Motroni, A. Buffi, P. Nepa, and B. Tellini, "Sensor-fusion and tracking method for indoor vehicles with low-density uhf-rfid tags," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.
- [9] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "Transvg: End-to-end visual grounding with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1769–1779.
- [10] J. Deng, Z. Yang, D. Liu, T. Chen, W. Zhou, Y. Zhang, H. Li, and W. Ouyang, "Transvg++: End-to-end visual grounding with language conditioned vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [11] F. Shi, R. Gao, W. Huang, and L. Wang, "Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [12] J. Ye, J. Tian, M. Yan, X. Yang, X. Wang, J. Zhang, L. He, and X. Lin, "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*, 2022, pp. 15502–15512.
- [13] L. Xiao, X. Yang, F. Peng, Y. Wang, and C. Xu, "Hivg: Hierarchical multimodal fine-grained modulation for visual grounding," *arXiv preprint arXiv:2404.13400*, 2024.
- [14] T. Liu, X. Liu, S. Huang, H. Chen, Q. Yin, L. Qin, D. Wang, and Y. Hu, "DARA: Domain- and relation-aware adapters make parameter-efficient tuning for visual grounding," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2024.
- [15] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19113–19122.
- [16] T. Liu, Y. Hu, W. Wu, Y. Wang, K. Xu, and Q. Yin, "Dap: Domain-aware prompt learning for vision-and-language navigation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.
- [17] L. Shi, B. Zhong, Q. Liang, N. Li, S. Zhang, and X. Li, "Explicit visual prompts for visual object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4838–4846.
- [18] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," in *Proceedings of the Advances in Neural Information Processing Systems*, 2022.
- [19] L. Xiao, X. Yang, F. Peng, M. Yan, Y. Wang, and C. Xu, "Clip-vg: Self-paced curriculum adapting of clip for visual grounding," *IEEE Transactions on Multimedia*, 2023.
- [20] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proceedings of the European Conference on Computer Vision*, 2016.
- [21] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [22] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Proceedings of the European Conference on Computer Vision*, 2016.
- [23] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.
- [24] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1307–1315.
- [25] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4683–4693.
- [26] M. Lu, R. Li, F. Feng, Z. Ma, and X. Wang, "Lgr-net: Language guided reasoning network for referring expression comprehension," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [27] K. Li, D. Wang, H. Xu, H. Zhong, and C. Wang, "Language-guided progressive attention for visual grounding in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [28] K. Li, F. Dong, D. Wang, S. Li, Q. Wang, X. Gao, and T.-S. Chua, "Show me what and where has changed? question answering and grounding for remote sensing change detection," *arXiv preprint arXiv:2410.23828*, 2024.
- [29] Y. Ding, H. Xu, D. Wang, K. Li, and Y. Tian, "Visual selection and multi-stage reasoning for rsvg," *IEEE Geoscience and Remote Sensing Letters*, 2024.
- [30] D. Liu, H. Zhang, F. Wu, and Z.-J. Zha, "Learning to assemble neural module tree networks for visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4673–4682.
- [31] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, "Learning to compose and reason with language tree structures for visual grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [32] Y. W. Chen, Y. H. Tsai, T. Wang, Y. Y. Lin, and M. H. Yang, "Referring expression object segmentation with caption-aware consistency," in *Proceedings of the British Machine Vision Conference*, 2019.
- [33] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2022.
- [34] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9499–9508.
- [35] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji, "Seqtr: A simple yet universal network for visual grounding," in *European Conference on Computer Vision*. Springer, 2022, pp. 598–615.
- [36] W. Su, P. Miao, H. Dou, Y. Fu, and X. Li, "Referring expression comprehension using language adaptive inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2357–2365.
- [37] H. Zhu, Q. Lu, L. Xue, M. Xue, G. Yuan, and B. Zhong, "Visual grounding with joint multi-modal representation and interaction," *IEEE Transactions on Instrumentation and Measurement*, 2023.
- [38] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," in *Proceedings of the International Conference on Learning Representations*, 2023.
- [39] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao et al., "Visionlm: Large language model is also an open-ended decoder for vision-centric tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [40] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song et al., "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.
- [41] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.
- [42] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "LoRA: Low-rank adaptation of large language models," in *Proceedings of the International Conference on Learning Representations*, 2022.
- [43] Y. Yuan, Y. Zhan, and Z. Xiong, "Parameter-efficient transfer learning for remote sensing image-text retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [44] X. Zhou, D. Liang, W. Xu, X. Zhu, Y. Xu, Z. Zou, and X. Bai, "Dynamic adapter meets prompt tuning: Parameter-efficient transfer learning for point cloud analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14707–14717.
- [45] Q. Wang, Y. Mao, J. Wang, H. Yu, S. Nie, S. Wang, F. Feng, L. Huang, X. Quan, Z. Xu et al., "Aprompt: Attention prompt tuning for efficient adaptation of pre-trained language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9147–9160.
- [46] T. Liu, X. Liu, L. Shi, Z. Xu, S. Huang, Y. Xin, and Q. Yin, "Sparse-Tuning: Adapting vision transformers with efficient fine-tuning and inference," *arXiv preprint arXiv:2405.14700*, 2024.
- [47] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

- [48] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [49] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai *et al.*, “Chatglm: A family of large language models from glm-130b to glm-4 all tools,” *arXiv preprint arXiv:2406.12793*, 2024.
- [50] T. Liu, Z. Xu, Y. Hu, L. Shi, Z. Wang, and Q. Yin, “Mapper: Multi-modal prior-guided parameter efficient tuning for referring expression comprehension,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 4984–4994.
- [51] X. Liu, T. Liu, S. Huang, Y. Hu, Q. Yin, D. Wang, and H. Chen, “M²ist: Multi-modal interactive side-tuning for memory-efficient referring expression comprehension,” *arXiv e-prints*, pp. arXiv–2407, 2024.
- [52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [53] M. Oquab, T. Dariseti, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khaldov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research*, 2023.
- [54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proceedings of the International Conference on Learning Representations*, 2020.
- [55] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [56] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
- [57] Z. Yang, T. Chen, L. Wang, and J. Luo, “Improving one-stage visual grounding by recursive sub-query construction,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 387–404.
- [58] M. Sun, W. Suo, P. Wang, Y. Zhang, and Q. Wu, “A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention,” *IEEE Transactions on Multimedia*, vol. 25, pp. 2446–2458, 2022.
- [59] H. Zhao, J. T. Zhou, and Y.-S. Ong, “Word2pix: Word to pixel cross-attention transformer in visual grounding,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 1523–1533, 2022.
- [60] C.-H. Ho, S. Appalaraju, B. Jasani, R. Manmatha, and N. Vasconcelos, “Yoro-lightweight end to end visual grounding,” in *European Conference on Computer Vision*. Springer, 2022, pp. 3–23.
- [61] A. J. Wang, P. Zhou, M. Z. Shou, and S. Yan, “Enhancing visual grounding in vision-language pre-training with position-guided text prompts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [62] P. Miao, W. Su, G. Wang, X. Li, and X. Li, “Self-paced multi-grained cross-modal interaction modeling for referring expression comprehension,” *IEEE Transactions on Image Processing*, 2023.
- [63] W. Tang, L. Li, X. Liu, L. Jin, J. Tang, and Z. Li, “Context disentangling and prototype inheriting for robust visual grounding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [64] W. Su, P. Miao, H. Dou, and X. Li, “Scanformer: Referring expression comprehension by iteratively scanning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 449–13 458.
- [65] H. Rezatofighi, N. Tsai, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.
- [66] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 69–85.
- [67] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [68] H. Lu, Y. Huo, G. Yang, Z. Lu, W. Zhan, M. Tomizuka, and M. Ding, “Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling,” in *Proceedings of the International Conference on Learning Representations*, 2024.
- [69] T. Liu, L. Shi, R. Hong, Y. Hu, Q. Yin, and L. Zhang, “Multi-stage vision token dropping: Towards efficient multimodal large language model,” *arXiv preprint arXiv:2411.10803*, 2024.