Look Around Before Locating: Considering Content and Structure Information for Visual Grounding

Shiyi Zheng[1], Peizhi Zhao[1], Zhilong Zheng[1], Peihang He[1], Haonan Cheng[2], Yi Cai[3], Qingbao Huang[1,4] *

[1]School of Electrical Engineering, Guangxi University, Nanning, China
[2]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China
[3]Key Laboratory of Big Data and Intelligent Robot of Ministry of Education, SCUT, Guangzhou, China
[4]Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning, China
qbhuang@gxu.edu.cn

Abstract

As a long-term challenge and fundamental requirement in vision and language tasks, visual grounding aims to localize a target referred by a natural language query. The regional annotations form a superficial correlation between the subject of expression and some common visual entities, which hinder models from comprehending the linguistic content and structure. However, current one-stage methods struggle to uniformly model the visual and linguistic structure due to the structural gap between continuous image patches and discrete text tokens. In this paper, we propose a semi-structured reasoning framework for visual grounding to gradually comprehend the linguistic content and structure. Specifically, we devise a cross-modal content alignment module to effectively align unlabeled contextual information into a stable semantic space corrected by token-level prior knowledge obtained with CLIP. A multi-branch modulated localization module is also established to obtain modulation grounding by linguistic structure. Through a soft split mechanism, our method can destructure the expression into a fixed semi-structure (i.e., subject and context) while ensuring the completeness of linguistic content. Our method is thus capable of building a semi-structured reasoning system to effectively comprehend the linguistic content and structure by content alignment and structure modulated grounding. Experimental results on five widely-used datasets validate the performance improvements of our proposed method.

Code — https://github.com/VILAN-Lab/SSRVG

Introduction

Visual grounding (VG) aims to comprehend the referring expression and locate the referred object in the image. As a fundamental task in the cross-modal domain, it attempts to build an entity-level link for a text-image pair, which facilitates numerous downstream studies, such as vision-and-language navigation (Anderson et al. 2018), image captioning (Chen et al. 2020), and visual question answering (Wang et al. 2020). Traditional paradigms, including early two-stage and one-stage methods (Liu et al. 2019b; Liao et al. 2020; Yang et al. 2020), focus on developing existing object detection frameworks. Typically, they utilize region features (Liu et al. 2019b) or point features (Liao et al. 2020; Yang et al. 2020) as the visual representation, and match or fuse with the linguistic embedding to rank or regress a bounding box. Some methods achieve meticulous modeling of linguistic information by modular design (Yu et al. 2018; Hu et al. 2017a) or graph mechanism (Yang, Li, and Yu 2020; Wang et al. 2019b), but they generally rely on numerous exact region proposals. Motivated by DETR (Carion et al. 2020), recent studies (Deng et al. 2021; Yang et al. 2022) attempt to establish a flexible single-step localization by the transformer framework which can effectively interact with multi-modal information. Concretely, (Deng et al. 2021; Ye et al. 2022) utilize a neat encoder for cross-modal information interaction, yet they struggle to align cross-modal information and localize the referred entity within a unified module. To tackle this issue, VLTVG assigns the alignment process to a shallow interactive encoder and modulates the localization process via a transformer decoder.

Although that recent works are fairly effective, the implicit modeling language structure by neural parameterizations may suffer from superficial bias (Cirik, Morency, and Berg-Kirkpatrick 2018; Akula et al. 2020), resulting in sub-optimal generalization. Concretely, for the case in Fig. 1,

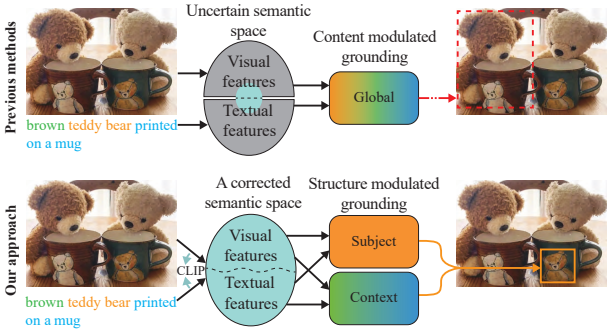

Figure 1: Interpretation of our designed idea. In previous methods, simple feature fusion leads to superficial predictions. In our method, a gradual process with fine-grained alignment corrected by external prior from CLIP modulates localization with semi-structural information.

*Corresponding author: Qingbao Huang

previous methods that rely on superficial content can lead to erroneous inferences in two ways: 1) Since annotations of contextual entities are not available at the training stage, the model can only learn a cross-modal feature space that is not completely aligned, leading the alignment process gradually to capture incorrect superficial correlations between common visual entities (e.g., bear toys) and the subject of the expression (e.g., *'Teddy Bear'*); 2)The uneven distribution of training data by category leads to the tendency of the model to directly select more common linguistic content as the referent target, instead of identifying the subject by exploring the structure of the expression.

Generally, a referring expression bundles a referent target and multiple context entities (Li and Jiang 2018), so the correct comprehension of the linguistic content and structure is the key to grounding. As shown in Fig. 1, to locate the referred entity (*"teddy bear"*) without ambiguity, it is necessary to exploit some structural information, such as attribute (*"brown"*) and relationship (*"printed on a mug"*). Thus, we consider designing a fine-grained alignment and localization process for visual grounding to sufficiently use the content and structure information of image and expression.

In this paper, we propose a semi-structured reasoning framework for visual grounding (SSRVG) which is a transformer-based method. As shown in Fig.1, different from previous methods, we design a gradual process to comprehend the expression content and structure from shallow to deep. To establish a fine-grained alignment for visual and linguistic contents, we first design a cross-modal content alignment (CCA) module that enables deep interaction between vision and language. Considering the alignment process may neglect the context due to the strong correlation between the subject and referred entity, we introduce token-level prior knowledge to build alignment for context in a corrected semantic space. Although explicit structured reasoning can be realized using graph mechanism (Yang, Li, and Yu 2020), it is difficult to fuse the continuous image patches and discrete expression structure in the one-stage framework due to the lack of consistent structure and grammatical rules (Qiao, Deng, and Wu 2021). Thus, we propose a soft split mechanism to parse the expression into semi-structures (i.e., subject and context). Based on the linguistic structure information, a multi-branch modulated localization (MML) module is designed to query the visual information by subject and context, respectively. By designing the MML module in cascade form, the problem of short information caused by the soft split mechanism can be alleviated. Our contributions can be summarized as follows: (1)To establish an effective fine-grained alignment in the case of insufficient annotations, we leverage the CLIP as a bridge to narrow the semantic gap. Our proposed CCA module can efficiently correct the alignment by token-level prior knowledge. (2)To the best of our knowledge, we are the first to modulate localization via semi-structured information of expressions. Our MML module with the cascade structure can alleviate the issue of overly short language segments caused by soft splitting. (3)The experiments demonstrate that our method outperforms conventional one-stage and transformer-based state-of-the-art methods on almost all benchmarks and it is comparable to some pre-trained models and large models.

Related Work

Two-stage methods (Hu et al. 2017b; Zhang, Niu, and Chang 2018; Wang et al. 2019b; Hong et al. 2022; Liu et al. 2019a; Wang et al. 2019a; Yang, Li, and Yu 2019; Zhuang et al. 2018; Chen et al. 2021) generally, achieve prediction by matching and ranking isolated visual regions with expressions. Earlier works (Yu et al. 2016; Mao et al. 2016) use the CNN-LSTM framework to model the multi-modal information. Module-based approaches (Yu et al. 2018; Hu et al. 2017a) are proposed to understand expression semantics more fully. MAttNet (Yu et al. 2018) implicitly splits subject, relationship, and location by attention mechanism. Some graph-based methods (Wang et al. 2019b; Yang, Li, and Yu 2020; Sun et al. 2023b) are proposed for the comprehensive modeling of structural information from complex text. Although they reduce cumulative error, two-stage approaches cannot focus on visual context because they only model isolated regions obtained by the object detector.

One-stage methods (Sadhu, Chen, and Nevatia 2019; Luo et al. 2020; Sun, Xiao, and Lim 2021; Sun et al. 2023a; Xiao et al. 2024; Su et al. 2024; Chen, Chen, and Wu 2024) usually pursue lightweight and celerity by a single-step localization based on point features. Most of them are designed based on a one-stage object detector, and apply a global representation of linguistic features to obtain a prediction. Concretely, (Yang et al. 2020; Li, Bu, and Cai 2021) attempt to exploit rich fine-grained information in expressions. (Yang et al. 2020) applies iterative sub-query to reduce long-term dependencies, while (Li, Bu, and Cai 2021) establishes a bottom-up multi-grained alignment. One-stage methods avoid accumulative errors, but the pre-defined dense anchors make prediction inflexible and accuracy questionable.

Notably, TransVG (Deng et al. 2021) achieves gratifying performance which avoids the limitation of pre-defined anchor boxes by a neat transformer framework. Most recent works (Yang et al. 2022; Du et al. 2022; Zhu et al. 2022; Ye et al. 2022) have followed this new **transformer-based** paradigm. To narrow the semantic gap between visual and language, (Yang et al. 2022) tries to constrain a shallow alignment by a verification module. Ye et al.(Ye et al. 2022) modulate the visual backbone by linguistic features to get the fusion representation. In addition, MDETR (Kamath et al. 2021), as a pre-trained model, extends DETR to numerous multi-modal tasks and uses the decoder to perform query-based set prediction, achieving more efficient localization. Benefiting from transformer architecture, transformer-based methods have advantages over conventional one-stage methods. However, these methods still have not established a reliable alignment and localization process. We try to solve these by semi-structured reasoning.

Approach

Overview

Given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a referring expression $L \in \mathbb{R}^{T}$, visual grounding aims to comprehend the natural language content and predict the bounding box $\hat{b} \in \mathbb{R}^{4}$ of the

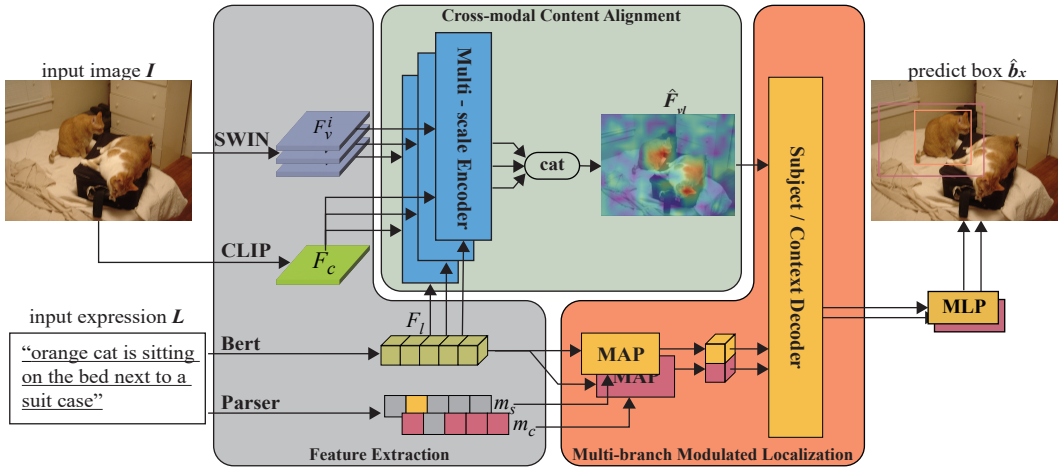

Figure 2: The overall framework of our **semi-structured reasoning visual grounding (SSRVG)** model. SWIN, CLIP, and BERT are used to extract features and Parser separates the subject and context. The encoder of our **cross-modal content alignment (CCA)** module is to build a fine-grained alignment for multi-scale visual and linguistic content by external prior. The decoder of our **multi-branch modulated localization (MML)** module utilizes semi-structured subject and context to modulate the localization.

referent. As shown in Fig. 2, our proposed semi-structured reasoning visual grounding (SSRVG) model contains three main components: backbones for feature extraction, a cross-modal content alignment (CCA) module as the encoder, and a multi-branch modulated localization (MML) module as the decoder. Concretely, the CCA module aligns vision and linguistic features to produce the multi-modal representation. Considering the scarcity of instance-level labels, we introduce CLIP as external knowledge to guide the alignment of cross-modal content. To utilize the complex text structure information sufficiently, the MML module decomposes expressions into subject and context, modulating information and locating the referred target in a cascade form.

Feature Extraction

Visual and linguistic branch. For the visual branch, we resize the image I to a fixed size and utilize a transformer network (i.e., Swin Transformer (Liu et al. 2021)) to extract the visual feature map $F_v \in \mathbb{R}^{H\times W\times d}$. For the linguistic branch, in addition to embedding the expression, we also extract partial structural information. Taking the input in Fig. 2 as an example, an expression includes a sequence of words $L = \{w_1, w_2, \ldots, w_T\}$. There is a subject word w_s in L, which is usually a noun denoting the class of the referent target or a pronoun. We use a dependency parser to locate the w_s and divide other words as context w_c. Then construct a subject mask $m_s \in \mathbb{R}^T$ and context mask $m_c \in \mathbb{R}^T$ according to w_s and w_c, while a sequence of textual tokens $F_l \in \mathbb{R}^{T\times d}$ is embedded by BERT (Devlin et al. 2019).

Semantic extraction branch. To leverage external knowledge to guide cross-modal alignment, we introduce a cross-modal semantic extraction branch. Specifically, we use the frozen visual branch of CLIP with the attention pooling layer removed to obtain a cross-modal representation $F_c \in \mathbb{R}^{H\times W\times d}$ from I. Considering the inconsistent tokenization between CLIP text encoder and BERT, we use the vision encoder to create semantic anchors F_c, utilized to guide the common visual and linguistic representations toward a reliable cross-modal semantic space narrowing the semantic gap between modalities on VG.

Cross-modal Content Alignment

CCA pipeline. To effectively align and fuse the initial uni-modal representation, multi-scale visual features $\{F_v^i\}_{i=3,4,5}$ are extracted by the Swin Transformer through the feature pyramid network and projected to the same dimension by convolutional blocks. CCA is then designed to extract a multi-modal representation from visual and linguistic content. To address the semantic gap between these two initial representations, which makes it challenging for the network to learn a fine-grained alignment for each instance, we propose a knowledge-guided cross-attention mechanism that narrows the gap by incorporating cross-modal prior knowledge. for each scale, the CCA module contains external prior correction and an M-layer transformer encoder.

CCA details. We first supplement the spatial context of F_v and the sequence context of F_l with a 2D-aware position embedding $E_v \in \mathbb{R}^{HW\times d}$ and 1D-aware position embedding $E_l \in \mathbb{R}^{T\times d}$, respectively. The multi-layer transformer encoder transfers information via the similarity metric for content, which is a common operation to align and fuse cross-modal representation by minimizing an optimization objective. However, different from the vanilla transformer encoder (Vaswani et al. 2017), we compute a correction matrix to ensure that semantic concepts lacking locate annotations can also be well aligned.

Corrected attention map. Concretely, we feed F_c into a self-attention (SA) module (the left one in Fig. 3 (a)) to obtain a patch token $\hat{F}_c \in \mathbb{R}^{HW\times d}$, feed F_l into a fully

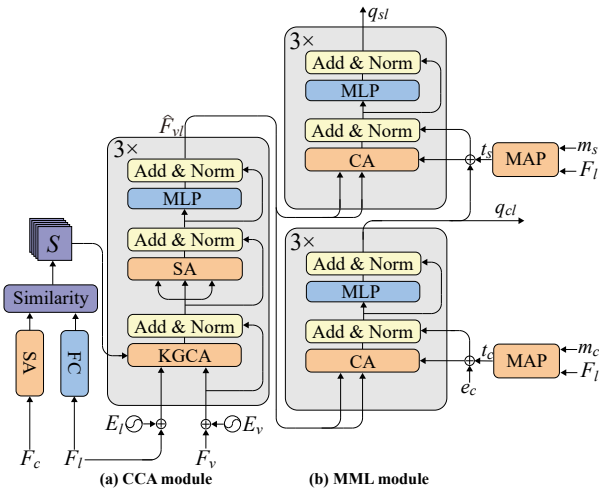

Figure 3: (a) Cross-modal content alignment (CCA) module, where **SA**, **KGCA** and **FC** are self-attention, knowledge-guided cross-attention and linear layer, respectively. (b) Multi-branch modulated localization (MML) module, where **MAP**, **CA**, and **MLP** are modulated attention pooling, cross-attention, and multi-layer perceptron, respectively.

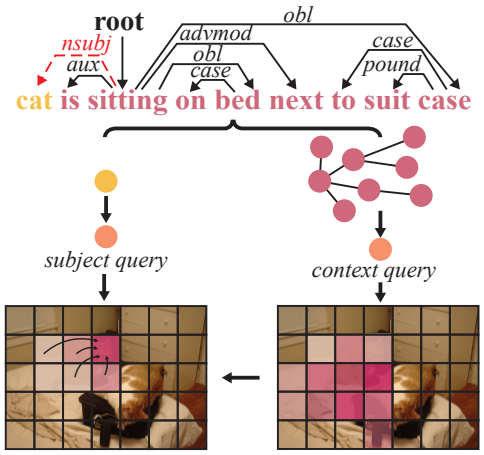

Figure 4: Interpretation of our semi-structured reasoning method. We first construct a semantic graph of discrete text tokens by dependency parser. Then a fixed relationship is used to decouple the graph into two parts. We pool the two sub-graphs into context-oriented nodes as the initial query and the subject-oriented nodes as the next stage query. For example, the context focuses on the suit case and the bed, while the subject focuses on the cat in the area ahead.

connected layer (FC) to project as $\hat{F}_l \in \mathbb{R}^{T\times d}$. A Gaussian kernel cosine similarity is exploited to measure a token-level affinity matrix $S \in \mathbb{R}^{HW\times T}$ which can be formulated as:

$$S = \alpha \cdot exp(-\frac{(1 - L2(\hat{F}_c)L2(\hat{F}_l)^T)^2}{2\sigma^2}), \quad (1)$$

where α and σ are learnable parameters, $L2$ is L2 normalization operator. To correct the alignment process, we design a knowledge-guided cross-attention (KGCA) layer before each encoder layer which can flexibly incorporate the affinity matrix into the attention map. The corrected attention map α_c can be calculated as:

$$Q = W_q(F_v^i + E_v), K = W_k(F_l + E_l) \quad (2)$$

$$\alpha_c = L1(\frac{S \odot exp(QK^T)}{\sqrt{d_k}}), \quad (3)$$

where Q and K are query and key, respectively, W_q and W_k are learnable embedding for the query and key, $\odot$ is Hadamard product, d_k is the channel dimension of the input, $L1$ is L1 normalization.

Other parts of CCA. After cross-modal interaction by the KGCA layer, a SA (the right one in Fig. 3 (a)) layer and a multi-layer perceptron (MLP) are utilized to further fuse the representations. The external representation vectors extracted from frozen CLIP, which can be viewed as a set of anchors on the cross-modal semantic space, are naturally closer to the raw linguistic or visual concepts. Compared to the dot-product attention, our corrected attention introduces the score S as an indirect optimization objective, which can force the model to focus on all text-related visual areas instead of being limited to the referred target. Finally, we take the output of the last block as the multi-modal representation $\hat{F}_{vl}$ which has a fine-grained alignment with the linguistic content through the CCA. The $\hat{F}_{vl}$ will be fed into the decoder which provides a better reference for localization in the later phases. The token-level information interaction aligns word features to corresponding visual patches indiscriminately, e.g., the words *'cat'* and *'bed'* are aligned with all corresponding entities in the image.

Multi-branch Modulated Localization

MML pipeline. Graph is usually used to represent structural information (Yang, Li, and Yu 2020), but one-stage VG methods encode images as spatially continuous visual patches, while texts are discrete tokens. This structural gap makes it impossible to model nodes and edges of graphs in the same semantic space. Therefore, we design a semi-structured reasoning framework that establishes a two-node graph through a fixed syntax. As shown in Fig. 4, the content of the classes and attributes used to describe the target is distributed in the subject and context of the expression, respectively. We separate the content into two sub-graphs according to the dependency relationship of *'nsubj'*, then use the global representation of the sub-graph as nodes which can avoid modeling of edges. Unlike cross-modal fusion decoder approaches (Deng et al. 2021; Ye et al. 2022), we propose a query-based decoder to model the association from textual node to image region. Similar to the query-based object detector (Carion et al. 2020), we attempt to establish a one-to-one mapping between node query and region to ensure the sparsity of the localization process.

MML details. As shown in Fig. 2, our MML module is a cascade form that can first perform rough contextual inference on semi-structured information, and then achieve more accurate localization based on the subject. A DETR decoder is utilized for MML to query the target in $\hat{F}_{vl}$ with the mod-

Models	Venue	Backbone	RefCOCO			RefCOCO+			RefCOCOg		ReferItGame	Flickr30K
			val	*testA*	*testB*	*val*	*testA*	*testB*	*val-u*	*test-u*		
fine-tuned model												
TransVG	ICCV'2021	ResNet-101	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	70.73	79.10
QRNet	CVPR'2022	Swin-S	84.01	85.85	82.34	72.94	76.17	63.81	73.03	72.52	74.61	81.95
VLTVG	CVPR'2022	Swin-S	84.69	87.54	82.32	74.28	79.22	67.95	74.86	75.11	70.26	80.57
Word2Pix	TNNLS'2022	ResNet-101	81.20	84.39	78.12	69.74	76.11	61.24	70.81	71.34	-	-
PLV-FPN	TIP'2022	ResNet-101	81.93	84.99	76.25	71.20	77.40	61.08	70.45	71.08	71.77	77.51
CLIP-VG	TMM'2023	ViT-B	84.29	87.76	78.43	69.55	77.33	57.62	73.18	72.54	70.89	81.99
TransVG++	TPAMI'2023	ViT-B	86.28	88.37	80.97	75.39	80.45	66.28	76.18	76.30	74.70	81.94
VG-LAW	CVPR'2023	ViT-B	86.06	88.56	82.87	75.74	80.32	66.69	75.31	75.95	**76.60**	-
PBREC	AAAI'2024	ResNet-101	82.94	86.31	80.81	74.85	79.53	65.60	73.86	74.13	-	-
LGR-NET	TCSVT'2024	Swin-S	85.63	88.24	82.69	75.32	80.60	68.30	76.82	77.03	74.64	81.97
EEVG	ECCV'2024	Swin-B	86.79	89.52	83.12	77.52	83.05	66.93	78.15	78.11	-	-
HiVG	ACMMM'2024	CLIP-B	87.32	89.86	83.27	78.06	84.81	68.11	78.29	78.79	75.22	82.11
SSRVG-T (ours)	-	ResNet-101	88.90	90.43	85.48	77.52	84.20	69.01	80.60	79.03	73.45	81.23
SSRVG-S (ours)	-	Swin-S	89.52	91.20	87.18	79.65	84.53	71.38	81.82	81.92	75.10	82.25
SSRVG-B (ours)	-	Swin-B	**90.45**	**91.72**	**88.04**	**80.14**	**84.56**	**72.57**	**82.45**	**82.88**	76.25	**83.01**
pre-trained model												
RefTR	NIPS'2021	ResNet-101	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01	-	-
MDETR	ICCV'2021	ResNet-101	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	-	-
UniTAB	ECCV'2022	ResNet-101	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	79.39	79.58
OFA	ICML'2022	OFA-B	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31	-	-
PolyFormer	CVPR'2023	Swin-B	89.73	91.73	86.03	83.73	88.60	76.38	84.46	84.96	**80.90**	-
LGR-NET	TCSVT'2024	Swin-S	88.16	90.01	84.26	78.66	83.51	73.18	82.60	82.58	77.78	82.18
EEVG	ECCV'2024	Swin-B	89.63	92.00	86.40	82.24	87.34	74.00	83.99	84.53	-	-
HiVG	ACMMM'2024	CLIP-B	90.56	92.55	87.23	83.08	89.21	76.68	84.52	85.62	77.75	82.08
HiVG-L	ACMMM'2024	CLIP-B	90.77	92.94	88.03	**86.78**	**89.91**	78.02	86.61	86.60	78.16	82.63
Multimodal Large Language Model												
Shikra-7B	-	-	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	-	-
Shikra-13B	-	-	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	-	-
Qwen-VL-7B	-	-	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48	-	-
Qwen-VL-7B-chat	-	-	88.55	92.27	84.51	82.82	88.59	76.79	85.96	86.32	-	-
SSRVG-B* (ours)	-	Swin-B	**92.73**	**94.04**	**90.75**	85.12	88.84	**78.97**	**88.30**	**88.51**	77.50	**83.15**

Table 1: Comparison with the state-of-the-art approaches on RefCOCO, RefCOCO+, RefCOCOg, ReferItGame, and Flickr30K in terms of top-1 accuracy (%). The best and second best performances are in **bold** and underlineunderline. * means model pre-trained.

ulated information. Considering that each image-text pair in the VG task only involves a specified singleton $\{C_L\}$ of online categories, we do not need to resort to a self-attention mechanism and assignment strategy to perform set prediction, which makes our decoder simpler and more efficient. Inspired by UP-DETR (Dai et al. 2021), we use visual features as queries for localization to the multi-modal domain, with the difference that our pre-defined queries are derived from expression subject or context.

MML output. As shown in Fig. 3 (b), to avoid the isolated information caused by separating the expression, a modulated attention pooling (MAP) module is designed to accomplish a soft split of F_l. Given the textual tokens F_l and mask m_x (i.e., m_s, m_c), we concatenate a modulated average pooling feature $f_l^x \in \mathbb{R}^d$ and F_l as $F_l^x = [f_l^x; F_l]$. The f_l^x is computed by $f_l^x = pool(m_x F_l)$, where $pool$ is an average pooling operator. As the query, F_l^x is fed into a self-attention layer to get $\hat{F}_l^x$. We choose the first vector in $\hat{F}_l^x$ as the modulated *query patch* $t_x \in \mathbb{R}^d$. We supply the m_s to get the t_s, and supply the m_c to get the t_c. Concretely, in the first three-layer of cross-attention, we sum t_c with a learnable *object query* $e_c \in \mathbb{R}^d$ as the initial region token $q_c^0 \in \mathbb{R}^d$. Then q_c^0 is fed into the attention layer which can iteratively extract the multi-modal features and updates itself into q_c^i ($0 \le i \le N$). In the *i*-th layer, the query q_c^{i-1} is fed into a cross-attention with the $\hat{F}_{vl}$ as the key and value. In the second three-layer of cross-attention, we sum t_s with the q_{cl} as the query, and the others are the same as the first. Finally, the $q_{sl}, q_{cl} \in \mathbb{R}^d$ are output by the decoder. For the prediction, we input q_{sl}, q_{cl} into a three-layer MLP and a Sigmoid layer to obtain the predicted bounding box $\hat{b}_s, \hat{b}_c \in \mathbb{R}^4$.

Optimization

Following DETR (Carion et al. 2020), a multi-layer prediction and auxiliary loss are used to speed up the convergence of the model. Specifically, for the output of each layer of the decoder, we calculate the sum $\hat{h}^i$, and use a shared prediction head to regress the bounding box of the target in the training stage. For the backward, we use the target box $\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$ and the ground-truth box $b = (x, y, w, h)$ to supervise the optimization of our model. We add a random jitter to the box, where the jitter value is a set percentage range and the cardinality is the original length and width of the box. The range of the ground truth for b_c is set to $[0, 0.05]$, while b_c is set to $[0.05, 0.15]$. Following Deng et al. (Deng et al. 2021), we use smooth L1 loss for coordinates and generalized IoU loss (Rezatofighi et al. 2019) for boxes as the optimization objectives:

$$\mathcal{L} = \sum\nolimits_{x \in \{c,s\}} \mathcal{L}_{smooth-l1}(b_x, \hat{b}_x) + \gamma \mathcal{L}_{giou}(b_x, \hat{b}_x), \quad (4)$$

where $\mathcal{L}_{smooth-l1}$ and $\mathcal{L}_{giou}$ are smooth L1 loss and GIoU loss, while γ is a hyperparameter to adjust the two losses.

	backbone	KGCA	multi-scale	structure	Acc (%)
1	Swin Base				79.06
2	Swin Base	✓			80.35
3	Swin Base		✓		79.80
4	Swin Base			✓	80.11
5	Swin Base	✓	✓		80.61
6	Swin Base	✓		✓	81.04
7	Swin Base		✓	✓	80.66
8	Swin Base	✓	✓	✓	82.45

Table 2: Ablation study of three components. multi-scale and structure represent the multi-scale feature for CCA module and the semi-structure reasoning for the MML module.

	Swin Base	BERT Base	CLIP Base	CLIP Base	
1	✓	✓			82.45
2	✓			✓	71.10
3		✓	✓		80.75
4			✓	✓	69.51

Table 3: Ablation study of text and vision encoders

	global	cont.	sub.	sub. & cont.	cont. & sub.	Acc%
1	✓					80.71
2		✓				81.51
3			✓			81.77
4				✓		82.10
5					✓	82.45

Table 4: Ablation study of Multi-branch modulated localization module. cont. means context and sub. means subject

Experiments

Datasets and Evaluation Metric

RefCOCO (Yu et al. 2016), **RefCOCO+** (Yu et al. 2016), and **RefCOCOg** (Mao et al. 2016) are the most widely used benchmarks for visual grounding, **ReferItGame** (Kazemzadeh et al. 2014) is sourced from SAIAPR-12 (Escalante et al. 2010). The annotations of ReferItGame are pervasively casual compared with other datasets, and numerous expressions are ungrammatical. **Flickr30K Entities** (Plummer et al. 2015) is essentially a phrase grounding dataset with 31,783 images and 427K referred objects. For a fair comparison, we evaluate our model by accuracy which is considered a correct prediction if the IoU between the predicted and ground truth is higher than 0.5. Each experiment is conducted three times, and the final results are averaged.

Implementation Details

For data preprocessing, se resize the image to a uniform size and pad the text sequence to a uniform length as 30 (e.g., RefCOCO, RefCOCO+, ReferItGame), 40 (e.g., RefCOCOg), 15 (Flickr30K Entities) or 40 for pre-training. To separate the subject of the expression, Stanza (Qi et al. 2020) is adopted to extract the subject. We follow (Carion et al. 2020; Deng et al. 2021; Yang et al. 2022) to augment the data. We use Swin Transformer Base (Swin Base) and ResNet-101 as the visual backbone for the network and use CLIP (Radford et al. 2021) encoder as an external plugin. During training, we use AdamW (Loshchilov and Hutter 2019) as the optimizer and set the initial learning rate to 1×10^{-4} for our network, except the visual backbone and BERT which have a lower initial learning rate of 10^{-6}. The weights of CLIP and Parser are frozen. For the number of encoder and decoder layers, we set M and N to 3 and 3, respectively. We train three versions of the model: SSRVG-T, SSRVG-S, and SSRVG-B with different visual backbones. SSRVG-B* is a pre-trained version of SSRVG-B. In all fine-tuned experiments, the learning rate scheduler is CosineAnnealingLR with the learning epochs set to 90. In pre-trained experiments, following previous works (e.g., HiVG(Xiao et al. 2024)), which used a mixed dataset pre-trained setting, we combine the training data from RefCOCO/+/g, ReferIt and Flickr30K for SSRVG-B*. The training is divided into two stages. In the first stage, we train with a fine-tuning setup on a mixed dataset. In the second stage, we freeze the text and visual encoders and fine-tune the specific datasets for 10 epochs with a learning rate 10^{-6}.

Comparison with State-of-the-art Models

To evaluate our method, we compare it with other state-of-the-art methods. As shown in Table 1, our model outperforms recent methods on almost all splits. Compared with HiVG (Lu et al. 2024), our model obtains improvements of up to 3.13%, 1.86%, 4.77%, 2.08%, -0.25%, 4.46%, 4.16%, 4.09%, 1.03% and 0.90%, respectively. Besides, our model surpasses the state-of-the-art two-stage method PBREC (Zhao et al. 2024) by 7.51%, 5.41%, 7.23%, 5.29%, 5.03%, 6.97%, 8.59%, 8.75% improvements on these five datasets. Compared with large generalist models, our fine-tuned model can achieve competitive results on RefCOCO, but lags behind large generalist models because of their larger parameters and richer training data on datasets (e.g. RefCOCO+, RefCOCOg) with more difficult referring expressions. On the ReferItGame, our model can achieve 76.25%, while it cannot obtain an absolute advantage compared with other models because the annotation quality of the dataset severely limits our method. So Stanza is unable to extract the subject from these low-quality descriptions effectively, which hinder the improvement of our model. The annotations of RefCOCO/+/g are more complex and complete, closer to the real scenarios, which makes many large models more inclined to evaluate these three datasets.

Ablation Study

All ablation studies are built on SSRVG-B and the performance improvements are verified on the RefCOCOg (valu). *The Component Modules:* We perform ablation experiments on the three main components, i.e., KGCA, multi-scale feature for cross-modal content alignment (CCA) module, and semi-structure for multi-branch modulated localization (MML) module. As shown in Table 2: 1) The first row shows our baseline that only utilizes a single-scale encoder without KGCA, which achieves an accuracy of 79.06%. 2) Adding KGCA to the baseline leads to an improvement by 1.29%. 3) Using multi-scale visual features brings 0.74% absolute improvements over the baseline. 4) Introducing semi-

structured reasoning over the baseline model significant improvements of 1.05%. 5) Using KGCA + multi-scale, the performance drops by 1.84% due to global pooling dispersing the expression semantic. 6) The absence of multiple scales result in a decrease of 1.41%. 7) Without CLIP guidance, the model perform fusion inference in an unknown space, resulting in a decrease of 1.79%.

Longer expression and complex scene composition: We divide the RefCOCOg test into three groups based on expression length and the results are that 1-5:89.29, 6-10:88.08, 11+:87.93. Besides, we also divide the RefCOCO testA into three groups based on instances and the result is that: 1-10:95.11, 11-20:92.87, 20+:91.46.

Textual and Vision Encoders: We have designed four combination experiments to verify the feasibility of the CLIP backbone (cf. Table 3). The model's performance using the textual encoder of CLIP decreases by 11.35% (Row1 vs Row2) and 11.24% (Row3 vs Row4) compared to BERT. Compared to BERT, which is pre-trained with Masked Language Modeling (MLM), CLIP focuses on the holistic representation of text. However, token-level features are more crucial in VG. While using the ResNet of CLIP, there is a slight decrease in accuracy because, although ResNet has powerful transfer learning capabilities, its performance is poorer than that of Swin Base. However, the consistent representation of images and text, as the most important capability of CLIP, will be affected when retraining on the VG task resulting in no improvement in accuracy.

Multi-branch modulated localization module: In Table 4, we provide five semantic environments for the decoder, i.e., global, subject, context, subject & context, and context & subject. The *global* means that the query patch comes from global attention pooling without modulation. The *subject* and *context* are normal query patches from modulated attention pooling. *subject & context* means subject-oriented nodes as the initial query and context-oriented nodes as the intermediate query, while *context & subject* is the opposite. Compared with only using global information, comprehensive utilization of subject and context can achieve a performance improvement of 1.74% because masking the subject compels the model to focus on contextual information for inference, thereby preventing it from over-relying on the subject during localization. There are still performance improvements of 0.80% (Row2 vs Row1) and 1.06% (Row3 vs Row1) when only using the context or subject as the main localization basis. Context-oriented nodes as the intermediate query bring a 0.35% improvement (Row5 vs Row4). The model searches for relevant regions in the image based on contextual information, and the subject helps to further refine this search. This form of interaction inspires the ability for progressive reasoning of model. The results show that effective use of the structural information of the expression by semi-structured reasoning can significantly improve the localization capabilities of our model.

Visualization

To illustrate more significantly the inhibitory effect of our semi-structured reasoning on superficial bias, we visualize the localization process for some difficult cases in Fig. 5.

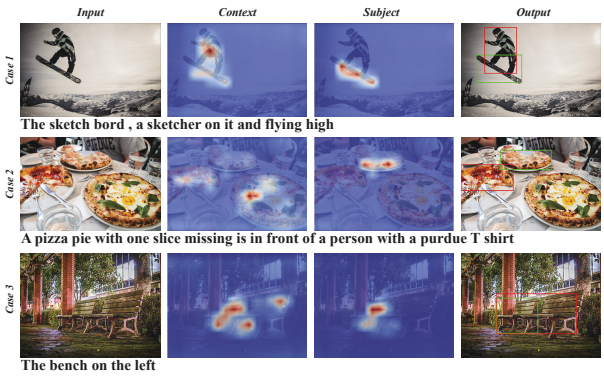

Figure 5: Visualization of the decoder's attention maps and the final prediction. The green box is the correct localization of our model, while the red box is the failed prediction of VLTVG. Context and subject are the attention map of the third and last layers of the decoder, respectively.

The red box is a failed prediction from VLTVG (Yang et al. 2022), and the green box is our correct localization. We visualize the attention maps of the decoder separately with heatmaps which can reflect the impact of different textual content on localization. Case 1 contains only one entity of target category, but due to the long-tail distribution of the dataset, VLTVG is more inclined to treat people as referred objects and thus fails to locate. Our decoder identifies two key descriptive objects: a human and a skateboard. Upon analyzing the input subject, the decoder confirms that the correct object is the skateboard. Furthermore, in the case that there are multiple entities of the same category as the referred object in the image, previous methods may not be able to focus on the correct entity by utilizing context information. In Case 2, our decoder is aware of text content such as *'one slice missing'* and *'in front of a person'*, which effectively corrects the localization process. Case 3 also shows that context can help the model understand visual information. Influenced by the *'left'* in the context, our model can significantly identify the two benches in the image. The results show that our semi-structured reasoning effectively balances the influence of subject and context by establishing the semi-structural expression.

Conclusions

In this paper, we propose SSRVG, a new transformer-based visual grounding method. To effectively comprehend the content and structure information of the expression, we establish a semi-structured reasoning process. Specifically, we build a fine-grained alignment for visual and linguistic content via correction of external prior and achieve a linguistic structure modulated localization by soft splitting the subject and context. Extensive experiments on public datasets show the advantages of our method. In the future, we plan to further investigate approaches to implement one-stage structured reasoning in the open domain for better generalization.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276072), Guangxi Natural Science Foundation Key Project (Application No. 2024JJD170001), and the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China (No. SKLMCC2023KF005), partly by the National Natural Science Foundation of China (62076100), the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology, the Science and Technology Planning Project of Guangdong Province (2020B0101100002), Guangdong Provincial Fund for Basic and Applied Basic Research-Regional Joint Fund Project (Key Project) (2023B1515120078), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), and Innovation Project of Guangxi Graduate Education (JGY2023016).

References

Akula, A. R.; Gella, S.; Al-Onaizan, Y.; Zhu, S.; and Reddy, S. 2020. Words Aren't Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions. In *ACL 2020*, 6555–6565.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I. D.; Gould, S.; and van den Hengel, A. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *CVPR 2018*, 3674–3683.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, 213–229.

Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S. 2021. Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding. In *AAAI 2021*, 1036–1044.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In *CVPR 2020*, 9959–9968.

Chen, W.; Chen, L.; and Wu, Y. 2024. An Efficient and Effective Transformer Decoder-Based Framework for Multi-Task Visual Grounding. *arXiv preprint arXiv:2408.01120*.

Cirik, V.; Morency, L.; and Berg-Kirkpatrick, T. 2018. Visual Referring Expression Recognition: What Do Systems Actually Learn? In *NAACL 2018*, 781–787.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. UP-DETR: Unsupervised Pre-Training for Object Detection With Transformers. In *CVPR 2021*, 1601–1610.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding with Transformers. In *ICCV 2021*, 1749–1759.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL 2019*, 4171–4186.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Visual Grounding with Transformers. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Escalante, H. J.; Hernández, C. A.; González, J. A.; López-López, A.; Montes-y-Gómez, M.; Morales, E. F.; Sucar, L. E.; Pineda, L. V.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *Comput. Vis. Image Underst.*, 114(4): 419–428.

Hong, R.; Liu, D.; Mo, X.; He, X.; and Zhang, H. 2022. Learning to Compose and Reason with Language Tree Structures for Visual Grounding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2): 684–696.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017a. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *CVPR 2017*, 4418–4427.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017b. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *CVPR 2017*, 4418–4427.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *IEEE/CVF International Conference on Computer Vision*, 1760–1770.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. L. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP 2014*, 787–798.

Li, L.; Bu, Y.; and Cai, Y. 2021. Bottom-Up and Bidirectional Alignment for Referring Expression Comprehension. In *ACM MM 2021*, 5167–5175.

Li, X.; and Jiang, S. 2018. Bundled Object Context for Referring Expressions. *IEEE Trans. Multim.*, 20(10): 2749–2760.

Liao, Y.; Liu, S.; Li, G.; Wang, F.; Chen, Y.; Qian, C.; and Li, B. 2020. A Real-Time Cross-Modality Correlation Filtering Method for Referring Expression Comprehension. In *CVPR 2020*, 10877–10886.

Liu, D.; Zhang, H.; Zha, Z.; and Wu, F. 2019a. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *ICCV 2019*, 4672–4681.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019b. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *CVPR 2019*, 1950–1959.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV 2021*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR 2019*.

Lu, M.; Li, R.; Feng, F.; Ma, Z.; and Wang, X. 2024. LGR-NET: Language Guided Reasoning Network for Referring Expression Comprehension. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *CVPR 2020*, 10031–10040.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR 2016*, 11–20.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *IEEE International Conference on Computer Vision*, 2641–2649.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 101–108.

Qiao, Y.; Deng, C.; and Wu, Q. 2021. Referring Expression Comprehension: A Survey of Methods and Datasets. *IEEE Trans. Multim.*, 23: 4426–4440.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 658–666.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-Shot Grounding of Objects From Natural Language Queries. In *ICCV 2019*, 4693–4702.

Su, W.; Miao, P.; Dou, H.; and Li, X. 2024. ScanFormer: Referring Expression Comprehension by Iteratively Scanning. arXiv:2406.18048.

Sun, M.; Suo, W.; Wang, P.; Zhang, Y.; and Wu, Q. 2023a. A Proposal-Free One-Stage Framework for Referring Expression Comprehension and Generation via Dense Cross-Attention. *IEEE Trans. Multim.*, 25: 2446–2458.

Sun, M.; Xiao, J.; and Lim, E. G. 2021. Iterative Shrinking for Referring Expression Grounding Using Deep Reinforcement Learning. In *CVPR 2021*, 14060–14069.

Sun, M.; Xiao, J.; Lim, E. G.; and Zhao, Y. 2023b. Cycle-Free Weakly Referring Expression Grounding With Self-Paced Learning. *IEEE Trans. Multim.*, 25: 1611–1621.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, 5998–6008.

Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2019a. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2): 394–407.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and van den Hengel, A. 2019b. Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks. In *CVPR 209*, 1960–1968.

Wang, X.; Liu, Y.; Shen, C.; Ng, C. C.; Luo, C.; Jin, L.; Chan, C. S.; van den Hengel, A.; and Wang, L. 2020. On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10123–10132.

Xiao, L.; Yang, X.; Peng, F.; Wang, Y.; and Xu, C. 2024. HiVG: Hierarchical Multimodal Fine-grained Modulation for Visual Grounding. arXiv:2404.13400.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9489–9498.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic Graph Attention for Referring Expression Comprehension. In *ICCV 2019*, 4643–4652.

Yang, S.; Li, G.; and Yu, Y. 2020. Graph-Structured Referring Expression Reasoning in the Wild. In *CVPR 2020*, 9949–9958.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving One-Stage Visual Grounding by Recursive Sub-query Construction. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12359, 387–404.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15481–15491.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9906 of *Lecture Notes in Computer Science*, 69–85.

Zhang, H.; Niu, Y.; and Chang, S. 2018. Grounding Referring Expressions in Images by Variational Context. In *CVPR 2018*, 4158–4166.

Zhao, P.; Zheng, S.; Zhao, W.; Xu, D.; Li, P.; Cai, Y.; and Huang, Q. 2024. Rethinking Two-Stage Referring Expression Comprehension: A Novel Grounding and Segmentation Method Modulated by Point. In *AAAI 2024*, 7487–7495.

Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; and Ji, R. 2022. SeqTR: A Simple Yet Universal Network for Visual Grounding. In *Computer Vision - ECCV 2022 - 17th European Conference*, volume 13695 of *Lecture Notes in Computer Science*, 598–615.

Zhuang, B.; Wu, Q.; Shen, C.; Reid, I. D.; and van den Hengel, A. 2018. Parallel Attention: A Unified Framework for Visual Object Discovery Through Dialogs and Queries. In *CVPR 2018*, 4252–4261.