Hierarchical Alignment-enhanced Adaptive Grounding Network for Generalized Referring Expression Comprehension

Yaxian Wang[1,2], Henghui Ding[3*], Shuting He[4], Xudong Jiang[5], Bifan Wei[6,7*], Jun Liu[1,2]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University, China
[3]Institute of Big Data, Fudan University, China
[4]Shanghai University of Finance and Economics, China
[5]Nanyang Technological University, Singapore
[6]School of Continuing Education, Xi'an Jiaotong University, China
[7]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, China
wyx1566@stu.xjtu.edu.cn, henghui.ding@gmail.com, shuting.he@sufe.edu.cn, {weibifan,liukeen}@xjtu.edu.cn

Abstract

In this work, we address the challenging task of Generalized Referring Expression Comprehension (GREC). Compared to the classic Referring Expression Comprehension (REC) that focuses on single-target expressions, GREC extends the scope to a more practical setting by further encompassing no-target and multi-target expressions. Existing REC methods face challenges in handling the complex cases encountered in GREC, primarily due to their fixed output and limitations in multi-modal representations. To address these issues, we propose a Hierarchical Alignment-enhanced Adaptive Grounding Network (HieA2G) for GREC, which can flexibly deal with various types of referring expressions. First, a Hierarchical Multi-modal Semantic Alignment (HMSA) module is proposed to incorporate three levels of alignments, including word-object, phrase-object, and text-image alignment. It enables hierarchical cross-modal interactions across multiple levels to achieve comprehensive and robust multi-modal understanding, greatly enhancing grounding ability for complex cases. Then, to address the varying number of target objects in GREC, we introduce an Adaptive Grounding Counter (AGC) to dynamically determine the number of output targets. Additionally, an auxiliary contrastive loss is employed in AGC to enhance object-counting ability by pulling in multi-modal features with the same counting and pushing away those with different counting. Extensive experimental results show that HieA2G achieves new state-of-the-art performance on the challenging GREC task and also the other 4 tasks, including REC, Phrase Grounding, Referring Expression Segmentation (RES), and Generalized Referring Expression Segmentation (GRES), demonstrating the remarkable superiority and generalizability of the proposed HieA2G.

Introduction

Generalized Referring Expression Comprehension (GREC) (He et al. 2023; Liu, Ding, and Jiang 2023) aims to detect an arbitrary number of target objects based on a given free-form text expression. In contrast to the classic Referring Expression Comprehension (REC) (Mao et al. 2016; Yu et al. 2016) that only supports the single-target text expressions, GREC narrows the gap with real-world scenarios by further encompassing no-target expressions that do not match any object in the image, and multi-target expressions that refer to multiple target objects. This task has great potential value for various practical applications such as visual-language navigation, embodied AI, and human-robot interaction.

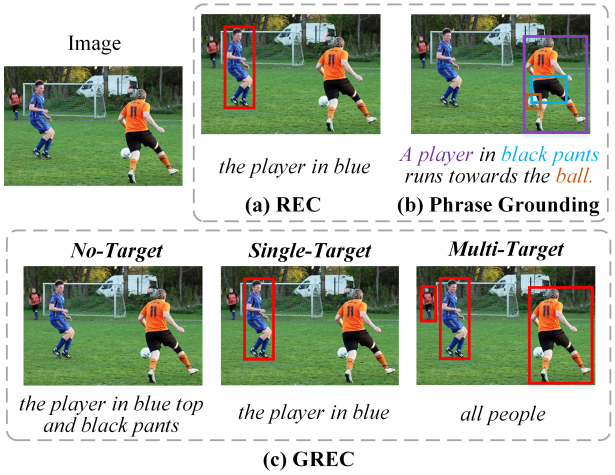

Figure 1: Different visual grounding tasks. (a) Classic REC: text expressions can only specify a single object; (b) Phrase grounding detects all objects mentioned in expressions; (c) GREC (He et al. 2023; Liu, Ding, and Jiang 2023) supports the text expressions indicating an arbitrary number of target objects from 0 to multiple, which is a more challenging task.

Although recent methods (Zhu et al. 2022; Deng et al. 2023) have achieved remarkable performance in REC, they are constrained to predict only one target that is most related to the text expression, leading to the inability to deal with no-target and multi-target expressions. As shown in Figure 1, the text expression "*the player in blue top and black pants*" does not match any object within the image. In this case, the traditional REC models (Kamath et al. 2021; Yan et al.

*Corresponding authors.

2023; Li and Sigal 2021; Luo et al. 2020) still produce a false-negative bounding box. When given the multi-target text expression "*all people*", existing REC models also fail to locate all matched targets in the image, arising from the fact that they are enforced to locate only a single target most related to the text expression. Despite phrase grounding (Plummer et al. 2015; Yu et al. 2020) can locate multiple objects, it tends to locate all objects based on the key noun phrases in the text expression without comprehending the entire text semantics. GREC is a more challenging task that requires a comprehensive understanding of the intricate semantics of text expressions and visual contents to handle any quantity of target objects ranging from zero to multiple. Therefore, it is necessary to advance a robust GREC model to adapt to this kind of complex generalized scenario.

The first challenge of GREC lies in how to effectively align the diverse text expressions with the corresponding images for comprehensive multi-modal understanding. Existing methods (Kamath et al. 2021; Radford et al. 2021; Liu, Jiang, and Ding 2024; Xu et al. 2022; Liu et al. 2022) have made significant efforts to alleviate the cross-modal semantic gap. Nevertheless, they tend to rely solely on single-level alignment, either word-object or text-image alignment, leading to insufficient vision-language interaction and further inhibiting the effective learning of exhaustive multi-modal information. For one image, diverse and flexible text expressions can specify different numbers of target objects from various perspectives as shown in Figure 1, highlighting the significance of multi-level cross-modal interactions. For example, for a no-target case, the model is often required to capture the fine-grained attribute details of local objects and have a comprehensive understanding of the global contextual information, further rejecting providing any object response. Therefore, it is far from satisfactory to handle the complex cases in GREC leveraging the single-level alignment between the flexible text expressions and images.

The second challenge of GREC is how to output different numbers of target objects dynamically for each specific image-text pair. Given a complex text expression specifying multiple targets, a potential approach is to split the text expression into multiple text expressions and query the model multiple rounds to obtain the target objects one by one. However, text expressions with implicit multi-target information are difficult to decompose, and such an approach can not solve the inherent requirement in GREC, which desires an efficient model to give all targets in a single forward process. More importantly, multi-target expressions such as "*three players*" and "*all people*" necessitate a model to possess an explicit or implicit object-counting ability. Although a threshold-based strategy (He et al. 2023) has demonstrated its advantages in selecting the target objects from multiple candidate object proposals, it is often challenging to decide an appropriate threshold for all samples. Moreover, using a unified threshold struggles to adapt to the characteristics of different samples, resulting in inaccurate prediction results for some samples. Therefore, it is crucial to design a more advanced strategy for selecting target objects.

To address the above challenges, we propose **HieA2G**, a **Hie**rarchical **A**lignment-enhanced **A**daptive **G**rounding Network, for GREC to deal with various types of referring expressions flexibly. Specifically, we design a Hierarchical Multi-modal Semantic Alignment (HMSA) module to achieve comprehensive and robust multi-modal understanding by coupling three levels of alignments including word-object, phrase-object, and text-image alignment. Due to the absence of fine-grained region-level annotations corresponding to the entity words, we propose a text mask recovery auxiliary task to reconstruct the masked text semantics with the visual object features to promote word-object alignment. In this way, the visual features are facilitated to fuse the semantic information of the masked entity fully. Compared with the entity word, the attribute-related information is essential to distinguish the object of the same category. By matching the descriptive phrase with the visual object, the encoder can derive distinctive visual features and understand a larger range of semantic units. After that, the high-level text-image alignment matches the overall semantics between the text and image, enabling a more comprehensive perception of global information. On the one hand, the HMSA module can help provide holistic and robust multi-modal understanding to facilitate more accurate object localization. On the other hand, it endows the model with the ability to exploit information at various levels of detail, which allows it to accomplish other various tasks like REC and phrase grounding. Furthermore, to address the varying number of target objects in GREC, we design an Adaptive Grounding Counter (AGC) to dynamically determine the number of output target objects for each specific image-text pair. Additionally, an auxiliary contrastive loss is employed in AGC to enhance the object-counting ability by pulling together the multi-modal features with the same counting and pushing away those with different counting.

Our contributions are summarized as follows: (1) We propose a Hierarchical Alignment-enhanced Adaptive Grounding Network (HieA2G) for GREC to support text expressions indicating an arbitrary number of target objects. (2) We design a Hierarchical Multi-modal Semantic Alignment module to enable hierarchical vision-language interactions across multiple levels for comprehensive and robust multi-modal semantic understanding. (3) We propose an Adaptive Grounding Counter to dynamically determine the number of output targets for each specific image-text pair, which can help deal with the multi/single/no-target text expressions flexibly. (4) Extensive experimental results show that HieA2G achieves new SOTA results on the challenging GREC task. It also exhibits superior performance across the other four visual grounding tasks including REC, Phrase Grounding, Referring Expression Segmentation (RES), and Generalized Referring Expression Segmentation (GRES).

Related Work

Referring Expression Comprehension. REC aims to detect one specific object from an image based on a referring expression. Existing methods can be classified into two groups: two-stage (Hu et al. 2017; Zhuang et al. 2018; Yang, Li, and Yu 2019; Liu et al. 2019a; Li, Bu, and Cai 2021) and one-stage (Liao et al. 2020; Zhou et al. 2021; Yang et al. 2022a; Deng et al. 2023; Ye et al. 2021) methods. The

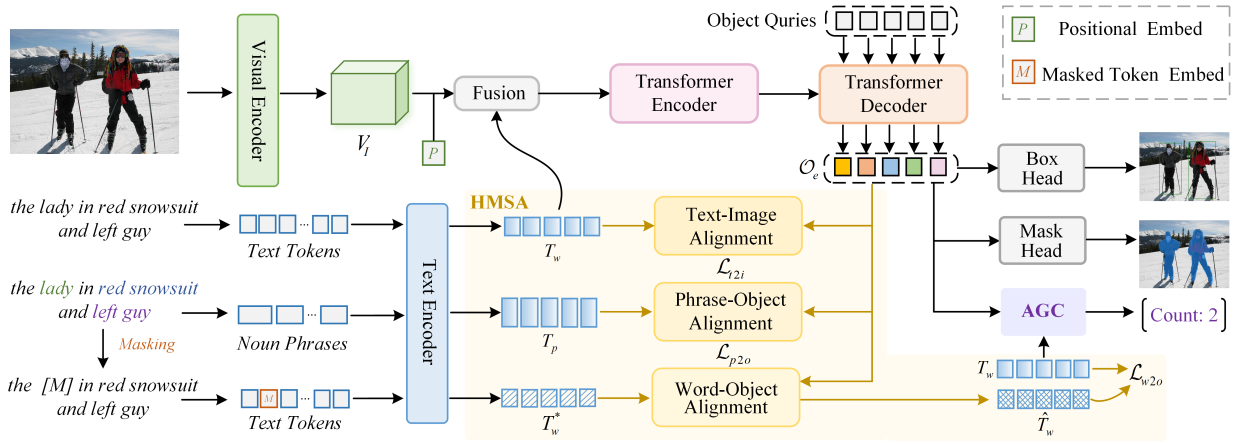

Figure 2: The framework of our proposed HieA2G. First, the visual encoder and the text encoder extract the visual feature V_I and text feature T_w. Then, a Transformer encoder is employed to perform multi-modal feature interaction further. The learnable object queries and the output of the Transformer encoder are fed to the Transformer decoder, whose output is object embeddings $\mathcal{O}_e$ corresponding to the object queries. Next, based on $\mathcal{O}_e$, the Hierarchical Multi-modal Semantic Alignment (HMSA) module is employed to facilitate multi-level cross-modal interaction via word-object, phrase-object, and text-image alignment. Moreover, an Adaptive Grounding Counter (AGC) is utilized to decide the output number of target objects dynamically.

recent advancements in large language models (Wu et al. 2024b; Chen et al. 2023; You et al. 2024) have also brought new opportunities to vision-language tasks requiring localization like REC. They have achieved promising results by collecting large-scale datasets, pretraining, and fine-tuning the large language models. It's worth noting that our work focuses on GREC, which detects an arbitrary number of target objects. Therefore, the top-1 selection method for REC cannot be applied directly in this setting. Although GREC (He et al. 2023) has modified some REC methods (Luo et al. 2020; Kamath et al. 2021; Ding et al. 2023c; Yan et al. 2023) to output different numbers of bounding boxes, they still struggle to deal with such complex and flexible referring expressions, leading to unsatisfactory performance.

Referring Expression Segmentation. RES aims to segment one object based on text expression. Driven by the success of Transformer, recent works (Li and Sigal 2021; Ding et al. 2021, 2023a,b,c; He and Ding 2024; He et al. 2024; Meng et al. 2022; Li et al. 2024; Wu et al. 2024a; Kim et al. 2022; Yang et al. 2022b; Wang et al. 2022) have extensively use it to extract visual and language features. ReLA (Liu, Ding, and Jiang 2023) introduces the Generalized Referring Expression Segmentation (GRES) benchmark, which further include multi-target and no-target samples. They only study region-language and region-image relationships. In contrast, we propose a hierarchical multi-modal alignment to enhance the comprehensive understanding of visual-linguistic context. Moreover, rather than a binary classification to judge only the existence of objects, we design an adaptive object-counting strategy to facilitate robust object perception.

Methodology

Architecture Overview

Figure 2 shows the overall architecture of our proposed HieA2G. First, the text expression T is fed into the text encoder to obtain the word features $T_w = \{w_k|k \in \{1, 2, ..., K\}\}$, where K is the number of words. For the phrase feature, we first extract noun phrases from the text expression, which are then represented as $T_p = \{p_i|i \in \{1, 2, ..., M\}\}$ by average pooling the word features in each phrase, where M is the number of phrases. For the image input I, we adopt a visual encoder to obtain the visual feature V_I and flatten it into a 2D feature combined with positional embeddings. The image feature and text feature are projected into the same space, and then are fused by concatenation to fed to the Transformer encoder for multi-modal deep fusion. Next, the output of the Transformer encoder and the learnable object queries are fed into the Transformer decoder. Subsequently, we obtain the text-aware object embeddings $\mathcal{O}_e = \{o_j|j \in \{1, 2, ..., N\}\}$ corresponding to the N object queries. Based on the object embeddings $\mathcal{O}_e$, the Hierarchical Multi-modal Semantic Alignment (HMSA) module is performed to hierarchically incorporate the multi-modal information across multiple levels for a comprehensive multi-modal understanding. Furthermore, we propose an Adaptive Grounding Counter (AGC) to determine the output number of target objects dynamically and then select the desired outputs from candidate object proposals.

Hierarchical Multi-modal Semantic Alignment

To fully model the relationships between the various types of text expression and the images, we propose a Hierarchical Multi-modal Semantic Alignment (HMSA) to facilitate multi-level cross-modal interactions through word-object, phrase-object, and text-image alignment. HMSA can help exploit information at various levels of detail and promote comprehensive and robust text-aware object embeddings for better box regression and mask segmentation.

Word-Object Alignment. To enable a more directly fine-grained word-object alignment, we introduce a masked text

recovery task by enforcing the model to recover the missed key information in the text based on the matched object features. The object embeddings are then facilitated to fuse the semantic information of the masked entity fully. Specifically, we first extract the entity noun in the text and randomly mask it with a [MASK] token. The masked text is encoded by the text encoder as T_w^*. Then, combining the object embeddings $\mathcal{O}_e$ and the masked text feature T_w^*, a Transformer layer is utilized to reconstruct the text semantic feature $\hat{T}_w$. We design a masked text recovery loss $\mathcal{L}_{w2o}$ by measuring the semantic similarity between the original complete text feature T_w and the reconstructed text feature $\hat{T}_w$ as:

$$\mathcal{L}_{w2o} = \alpha(1 - \cos(T_w, \hat{T}_w)), \tag{1}$$

where α is set to 0 when given a no-target sample, otherwise set to 1. Due to the weak relevance and even total irrelevance between the textual and the visual features for no-target samples, it is meaningless and impossible to reconstruct the missing information for this kind of image-text pair, even interfering with model optimization.

Phrase-Object Alignment. With the explicit phrase-object annotations in the Flickr30K Entities (Plummer et al. 2015) dataset, it is promising to encourage the model to match each phrase with the corresponding object query. By matching the descriptive phrase with the visual object, we can derive distinctive object features and understand a larger range of semantic units. Specifically, we first project the phrase features $T_p \in \mathbb{R}^{M \times C_p}$ and the object embeddings $\mathcal{O}_e \in \mathbb{R}^{N \times C_v}$ into the same sub-space by linear layers:

$$\widehat{T}_p = W_1 T_p, \quad \widehat{\mathcal{O}}_e = W_2 \mathcal{O}_e, \tag{2}$$

where W_1 and W_2 are learnable parameters, $\widehat{T}_p \in \mathbb{R}^{M \times C}$ and $\widehat{\mathcal{O}}_e \in \mathbb{R}^{N \times C}$ are the projected phrase features and projected object embeddings, M and N is the number of phrases and object queries, respectively. The matching relation map $S \in \mathbb{R}^{M \times N}$ between all noun phrases and the object queries is calculated as follows:

$$S = \text{Sigmoid}(\widehat{T}_p \cdot \widehat{\mathcal{O}}_e^{\top}). \tag{3}$$

For all object queries, we adopt a bipartite matching (Cheng et al. 2022) to find the matched ground-truth bounding box. Then, we can obtain a ground-truth binary map $Y \in \mathbb{R}^{M \times N}$ between phrases and object queries, indicating their matching relationships. With the predicted matching relation map $S \in \mathbb{R}^{M \times N}$, we design a phrase-object contrastive loss $\mathcal{L}_{p2o}$, implemented with the binary cross-entropy loss:

$$\mathcal{L}_{p2o} = -\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{i,j}\log S_{i,j} + (1-Y_{i,j})\log(1-S_{i,j}). \tag{4}$$

In this way, the model is encouraged to generate higher scores for positive phrase-object alignments and lower scores for negative ones. Therefore, the object embeddings can be endowed with stronger discriminative ability by capturing fine-grained attributes within the phrase semantics.

Text-Image Alignment. The image feature should have high feature similarity with the matched text expression and low feature similarity with the unmatched text expression. To fully model the global relationships between the image I and text expression T, we define a global match score $S^T(I,T)$ for each image-text pair to calculate their similarity via word-object pairs as follows:

$$S^T(I,T) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} a_{j,k}\langle \hat{o}_j, \hat{w}_k \rangle, \tag{5}$$

$$a_{j,k} = \frac{\exp\langle \hat{o}_j, \hat{w}_k \rangle}{\sum_{l=1}^{K} \exp\langle \hat{o}_j, \hat{w}_l \rangle}, \tag{6}$$

where $\langle .,. \rangle$ represents the dot product operation of two embeddings, $\hat{o}_j$ and $\hat{w}_k$ denotes the projected embedding of j-th object query and k-th word, and $S^T(I,T)$ is computed by normalizing along the text dimension. Similarly, $S^I(I,T)$ can be obtained by normalizing along the image dimension.

The global match score $S^T(I,T)$ measures the degree of semantic alignment between an image and its corresponding text expression. In this case, maximizing the match score of matched image-text pairs helps ensure their strong correspondence. For an image-text pair in a batch, the objective function is defined as follows:

$$\mathcal{L}_{t2i}^{TT}(I) = -\log \frac{\exp(S^T(I,T))}{\sum_{T' \in \mathcal{B}_T} \exp(S^T(I,T'))}, \tag{7}$$

$$\mathcal{L}_{t2i}^{TI}(T) = -\log \frac{\exp(S^T(I,T))}{\sum_{I' \in \mathcal{B}_I} \exp(S^T(I',T))}, \tag{8}$$

where $\mathcal{B}_T$, $\mathcal{B}_I$ represents a collection of the text expressions and images in a batch, $\mathcal{L}_{t2i}^{TT}(I)$ and $\mathcal{L}_{t2i}^{TI}(T)$ are normalized along text and image dimension, respectively. Similarly, $\mathcal{L}_{t2i}^{IT}(I)$ and $\mathcal{L}_{t2i}^{II}(T)$ can be obtained using $S^I(I,T)$.

The final text-image alignment loss for each image-text pair is computed as follows:

$$\mathcal{L}_{t2i} = \mathcal{L}_{t2i}^{TT}(I) + \mathcal{L}_{t2i}^{IT}(I)) + \mathcal{L}_{t2i}^{TI}(T) + \mathcal{L}_{t2i}^{II}(T), \tag{9}$$

In this way, the thorough text-image alignment boosts a more comprehensive understanding of global multi-modal information.

The final loss of HMSA is computed as $\mathcal{L}_{align} = \mathcal{L}_{w2o} + \mathcal{L}_{p2o} + \mathcal{L}_{t2i}$. By incorporating these three-level alignments, the object embeddings corresponding to the object queries can be gradually refined for better box regression and mask segmentation, further obtaining N potential object proposals via box head and mask head.

Adaptive Grounding Counter

To adapt to the generalized setting such as no-target and multi-target samples, we design an Adaptive Grounding Counter (AGC) to decide the output number of target objects dynamically for each specific image-text pair. With AGC, the desired target objects can be selected effectively from the N candidate object proposals corresponding to the object queries. Specifically, we formulate it as a classification task to predict an output number. For different images, the same text expression can specify different target objects. For

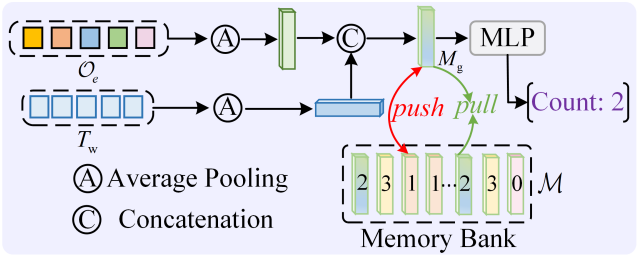

Figure 3: The detail of the Adaptive Grounding Counter.

example, "*all kids*" can denote an arbitrary number of objects in different images. Therefore, it is necessary to incorporate both the sentence feature of the text expression and object embeddings to predict the label. To better train AGC, we conducted a statistical analysis of all samples in the gRefCOCO dataset (Liu, Ding, and Jiang 2023; He et al. 2023). The distribution of objects follows a long-tailed pattern, with most samples falling within the range of 0 to 3, and only a small proportion exceeding the number of 3.

Therefore, it is defined as a classification task into five classes. With word features T_w and object embeddings $\mathcal{O}_e$ at hand, we adopt the average pooling to obtain the global text feature and visual feature. Then, the two features are concatenated to obtain a global multi-modal feature M_g, which is used to predict the object counting label y_c as:

$$M_g = [\text{AP}(T_w); \text{AP}(\mathcal{O}_e)], \tag{10}$$

$$y_c = \text{MLP}(M_g), \tag{11}$$

where [;] denotes concatenation operation, AP denotes average pooling, MLP denotes a two-layer perceptron, and $y_c \in \{0, 1, 2, 3, 3+\}$. Note that, only when the counting is larger than 3, the threshold-based strategy is adopted, that is object proposals with class scores above the threshold are selected. Otherwise, the sorted target objects with high scores are selected according to the counting.

To further promote the object counting ability, we incorporate contrastive learning in AGC by pulling together the multi-modal features M_g with the same counting and pushing away those with the different counting in Figure 3. The number of negative samples is related to batch size. However, the size of the batch size is limited by GPU memory. Therefore, to facilitate contrastive learning, we introduce a memory bank $\mathcal{M}$ (He et al. 2020) to maintain a larger number of negative samples. Inspired by (Khosla et al. 2020), a supervised contrastive loss $\mathcal{L}_{con}$ is introduced as follows:

$$\mathcal{L}_{con} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(M_g^i \cdot M_g^p / \tau)}{\sum_{a \in A(i)} \exp(M_g^i \cdot M_g^a / \tau)}, \tag{12}$$

where i denotes anchor index, $P(i) = \{p \in A(i), y_c^p = y_c^i\}$ is a collection of indices for the positive samples in $\mathcal{M}$, $|P(i)|$ denotes the cardinality of the collection, $A(i)$ denotes a collection of indices for all positive and negative samples in $\mathcal{M}$, M_g denotes the global multi-modal feature, and τ is a temperature hyperparameter.

The final loss of AGC is computed as $\mathcal{L}_{agc} = \mathcal{L}_{cls} + \mathcal{L}_{con}$, where $\mathcal{L}_{cls}$ denotes the object counting classification loss, implemented by a cross-entropy loss.

Methods	val		testA		testB	
	Pr	N-acc.	Pr	N-acc.	Pr	N-acc.
MCN†	28.0	30.6	32.3	32.0	26.8	30.3
VLT†	36.6	35.2	40.2	34.1	30.2	32.5
MDETR†	42.7	36.3	50.0	34.5	36.5	31.0
UNITEXT†	58.2	50.6	46.4	49.3	42.9	48.2
Ferret*	54.8	48.9	49.5	45.2	43.5	43.8
HieA2G$_{\text{R101}}$	**67.8**	**60.3**	**66.0**	**60.1**	**56.5**	**56.0**

Table 1: Results on gRefCOCO dataset (Liu, Ding, and Jiang 2023) in terms of Pr@(F_1=1, IoU≥0.5) and N-acc. for GREC task. † denotes these methods have been modified to generate multiple boxes following (He et al. 2023). $*$ denotes the model adapted for the GREC task.

Training Objective

To further supervise task-specific training, a series of losses for the box head and mask head are introduced as follows:

$$\mathcal{L}_{det} = \lambda_{bbox}\mathcal{L}_{bbox} + \lambda_{giou}\mathcal{L}_{giou} + \lambda_{class}\mathcal{L}_{class}, \tag{13}$$

$$\mathcal{L}_{seg} = \lambda_{mask}\mathcal{L}_{mask} + \lambda_{dice}\mathcal{L}_{dice}, \tag{14}$$

where λ_* are the hyperparameters, $\mathcal{L}_{class}$ is cross-entropy loss for box classification, $\mathcal{L}_{bbox}$ and $\mathcal{L}_{giou}$ are L1 loss (Ren et al. 2015) and GIoU loss (Rezatofighi et al. 2019) for box regression. The focal loss $\mathcal{L}_{mask}$ (Lin et al. 2017) and dice loss $\mathcal{L}_{dice}$ (Milletari, Navab, and Ahmadi 2016) are introduced to supervise mask segmentation (Ding et al. 2018).

We first pretrain HieA2G on a combined dataset formed by the training data of RefCOCO/+/g, Flickr30K Entities, and gRefCOCO datasets using the joint loss $\mathcal{L}_{pretrain} = \mathcal{L}_{align} + \mathcal{L}_{det}$. The goal of pretraining is to incorporate comprehensive multi-modal information into object queries. Then based on the pretrained weights, HieA2G is finetuned on various downstream tasks with the task-specific loss such as $\mathcal{L}_{det}$, $\mathcal{L}_{mask}$ and an additional loss $\mathcal{L}_{agc}$ introduced specifically for GREC and GRES.

Experiments

Experimental Setup

Datasets. The proposed HieA2G is evaluated mainly on the gRefCOCO dataset (He et al. 2023; Liu, Ding, and Jiang 2023) for GREC and GRES. We also conducted experiments on a phrase grounding dataset called Flickr30K Entities (Plummer et al. 2015), and three widely-used REC and RES benchmarks including RefCOCO (Yu et al. 2016), RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016).
Implementation Details. We adopt ResNet101 (He et al. 2016) and Swin-B (Liu et al. 2021) as our visual encoder, and RoBERTa-base (Liu et al. 2019b) as our text encoder.

Performance Comparison

Results on GREC. As shown in Table 1, our HieA2G with the ResNet101 backbone achieves superior performance on both metrics across three splits of the gRefCOCO dataset. It shows an average performance gain of 14.2% in Pr@(F_1=1, IoU ≥ 0.5) over Ferret (You et al. 2024) using a Multimodal

Figure 4: Visualization for the success cases and failure cases of HieA2G on gRefCOCO dataset. The ground truth is denoted by red bounding boxes, whereas green bounding boxes denote the predictions. The F_1 score of all success cases in (A) is 1.0.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
MAttNet	76.7	81.1	70.0	65.3	71.6	56.0	66.6	67.3
RefTR	85.7	88.7	81.2	77.6	82.3	69.0	79.3	80.0
MDETR	86.8	89.9	81.4	79.5	84.1	70.6	81.6	80.9
SeqTR	83.7	86.5	81.2	71.5	76.3	64.9	74.9	74.2
TransVG++	86.3	88.4	81.0	75.4	80.5	66.3	76.2	76.3
LISA-7B	85.4	88.8	82.6	74.2	79.5	68.4	79.3	80.4
GSVA-7B	86.3	89.2	83.8	72.8	78.8	68.0	81.6	81.8
HieA2G$_{R101}$	**87.8**	**90.3**	**84.0**	**80.7**	**85.6**	**72.9**	**83.7**	**83.8**

Table 2: Results comparison on RefCOCO/+/g for REC task.

Methods	val			test		
	R@1	R@5	R@10	R@1	R@5	R@10
VisualBert	68.1	84.0	86.2	-	-	-
VisualBert	70.4	84.5	86.3	71.3	85.0	86.5
MDETR	82.5	92.9	94.9	83.4	93.5	95.3
Shrika-7B	75.8	-	-	76.5	-	-
Ferret-7B	80.4	-	-	82.2	-	-
HieA2G$_{R101}$	**82.9**	**93.2**	**95.1**	**83.7**	**93.8**	**95.6**

Table 3: Results comparison on Flickr30K Entities dataset in Recall@k (ANY-BOX protocol) for Phrase Grounding task.

Large Language Model (MLLM), and an average performance gain of 9.4% in N-acc. over UNITEXT (Yan et al. 2023). These results indicate that HieA2G has a significant advantage in handling various types of text expressions to flexibly detect target objects ranging from zero to multiple.

Results on REC. As illustrated in Table 2, HieA2G achieves consistent performance gains across all splits of the three datasets compared to existing classic REC methods. HieA2G with the ResNet101 backbone even outperforms GSVA-7B (Xia et al. 2024) based on MLLM. The promising results can be attributed to the hierarchical multi-modal semantic alignment design, which promotes a comprehensive understanding of information at different granularities.

Results on Phrase Grounding. The main results on the

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
MAttNet	56.5	62.4	51.7	46.7	52.4	40.1	47.6	48.6
MCN	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT	65.7	68.3	62.7	55.5	59.2	49.4	52.9	56.7
HieA2G$_{R101}$	**73.3**	**75.9**	**69.0**	**64.8**	**69.7**	**56.1**	**62.9**	**63.5**
LAVT	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA-7B	74.9	**79.1**	72.3	65.1	70.8	58.1	67.9	70.6
GSVA-7B	**77.2**	78.9	**73.5**	65.9	69.6	**59.8**	**72.7**	**73.3**
HieA2G$_{SwinB}$	75.1	77.6	71.1	**66.5**	**71.4**	58.9	65.3	66.6

Table 4: Results comparison on RefCOCO/+/g for RES task.

Flickr30K Entities are shown in Table 3. HieA2G with ResNet101 improves performance over the previous SOTA MDETR on both val and test splits, suggesting that our method effectively enhances the multi-modal interactions.

Results on RES. As shown in Table 4, HieA2G outperforms the previous SOTA method ReLA (Liu, Ding, and Jiang 2023) with the same Swin-B backbone. It also shows competitive performance on RefCOCO and RefCOCO+ datasets to MLLM-based LISA-7B (Lai et al. 2024) and GSVA-7B. The results demonstrate that the comprehensive multi-modal representation ability of our HieA2G can contribute a lot to accurate segmentation for referring objects.

Results on GRES. GRES aims to generate masks for an arbitrary number of target objects. Unlike the previous SOTA GRES method ReLA using a simple binary classification branch for object-existence judgment, HieA2G has an explicit object-counting ability to facilitate accurate object perception in the generalized scenario. In Table 5, HieA2G with the Swin-B backbone achieves clear performance improvements over ReLA across all three evaluation sets on different metrics. It is even slightly better than the strong MLLM-based GSVA-7B in CIoU and GIoU. Besides generating high-quality masks, HieA2G with either backbone demonstrates outstanding performance in N-acc. and T-acc.,

Methods	Backbone	val				testA				testB			
		cIoU	gIoU	N-acc.	T-acc.	cIoU	gIoU	N-acc.	T-acc.	cIoU	gIoU	N-acc.	T-acc.
MAttNet	ResNet101	47.5	48.2	41.2	96.1	58.7	59.3	44.0	97.6	45.3	46.1	41.3	95.3
VLT	DarkNet53	52.5	52.0	47.2	95.7	62.2	63.2	48.7	95.9	50.5	50.9	47.8	94.7
VLT+ReLA	DarkNet53	58.7	59.4	-	-	66.6	65.4	-	-	56.2	57.4	-	-
CRIS	ResNet101	55.3	56.3	-	-	63.8	63.4	-	-	51.0	51.8	-	-
HieA2G	ResNet101	**62.5**	**67.1**	**60.9**	**97.4**	**67.6**	**70.5**	**60.2**	**97.7**	**58.8**	**61.5**	**56.5**	**96.4**
LAVT	Swin-B	57.6	58.4	49.3	96.2	65.3	65.9	49.3	95.1	55.0	55.8	48.5	95.3
ReLA	Swin-B	62.4	63.6	56.4	96.3	69.3	70.0	59.0	97.8	59.9	61.0	58.4	95.4
LISA-7B	ViT-H	61.8	61.6	54.7	-	68.5	66.3	50.0	-	60.6	58.8	51.9	-
GSVA-7B	ViT-H	63.3	66.5	62.4	-	69.9	71.1	**65.3**	-	60.5	62.2	60.6	-
HieA2G	Swin-B	**64.2**	**68.4**	**62.8**	**98.3**	**70.4**	**72.0**	63.4	**98.5**	**61.0**	**62.8**	**60.8**	**97.5**

Table 5: Results comparison on gRefCOCO dataset in terms of cIoU, gIoU, N-acc. and T-acc. for GRES task.

#	HMSA			AGC		GREC	
	W2O	P2O	T2I	Classifier	$\mathcal{L}_{con}$	Pr	N-acc.
#1				✓	✓	65.2	54.9
#2		✓	✓	✓	✓	67.5	58.1
#3	✓		✓	✓	✓	67.1	57.3
#4	✓	✓		✓	✓	67.0	56.4
#5	✓	✓	✓			53.9	48.0
#6	✓	✓	✓	✓		66.5	57.3
#7	✓	✓	✓	✓	✓	**67.8**	**60.3**

Table 6: Ablation study of different components of HieA2G for GREC. Notably, W2O, P2O and T2I indicate the word-object, phrase-object, and text-image alignment.

indicating its robust object perception ability.

Ablation Study

Effect of Hierarchical Multi-modal Semantic Alignment. From the first to fourth rows of Table 6, we perform an in-depth study of the HMSA module to validate its effectiveness. In the first row, removing all three-level alignments of the HMSA module leads to a significant decrease of 2.6% in Pr@(F_1=1, IoU $\geq$ 0.5) and 5.4% in N-acc. Furthermore, we can find that removing any one of the alignments, including W2O, P2O, and T2I, all leads to performance degradation compared with the overall model in the seventh row. This demonstrates that different levels of alignment can refine the object embeddings of the object queries, further facilitating accurate grounding by combining them.

Effect of Adaptive Grounding Counter. In the last three rows of Table 6, we test the effectiveness of AGC. The overall AGC is removed in the fifth row and replaced by the default threshold-based strategy (He et al. 2023) to filter the output objects. We can observe a 13.9% and 12.3% performance drop in terms of Pr@(F_1=1, IoU $\geq$ 0.5) and N-acc., which reflects that our advanced adaptive selection strategy contributes a lot to the output of target objects. Then, we add the classifier in the sixth row, which achieves 12.6% and 9.3% performance gain in both metrics respectively. When combined with $\mathcal{L}_{con}$ in the last row, further improvement can be brought for all metrics. This suggests that $\mathcal{L}_{con}$ is helpful to enhance the model's object counting ability.

Qualitative Analysis

We visualize some qualitative examples of our method on the validation split of gRefCOCO dataset to discuss the strengths and weaknesses of HieA2G as shown in Figure 4.

Analysis of Success Cases. As shown in (A) of Figure 4, our model can deal with various complex multi-target expressions in (a)-(d) and no-target expressions in (e)-(f). For example, HieA2G can count accurately "*two airplanes*" in (a) with shared attributes, and can differentiate an ordinal number like "2nd" to detect the correct bus in (b). It can also explicitly recognize all specified target objects for complex text expressions in (b), (c), and (d). Moreover, HieA2G can grasp the fine-grained attribute details to reject the no-target expression "*the board in man hand*" in (e). It has a comprehensive understanding of the global contextual information of all objects in (f), thereby rejecting to give a detection result due to no object in the image satisfying the description.

Analysis of Failure Cases. We show some failure cases of HieA2G in (B) of Figure 4. There are two main types of failure cases. ❶ For the three cases (a)-(c) in the first row, due to the ambiguous visual clues in the image, HieA2G struggles to detect all target objects for the first two cases and fails to reject giving a target for the last case. ❷ For the three cases (d)-(f) in the second row, due to the occlusion of the key visual clues, HieA2G fails to detect a desired object for the first two cases and gives a false negative target for the last no-target case. The analysis of failure cases reveals the limitations of HieA2G, while also shedding light on potential directions for our future research.

Conclusion

We propose a Hierarchical Alignment-enhanced Adaptive Grounding Network (HieA2G) for the challenging GREC task. The proposed Hierarchical Multi-modal Semantic Alignment (HMSA) module enables multi-level cross-modal interactions to achieve comprehensive and robust multi-modal understanding for better grounding. Adaptive Grounding Counter (AGC) determines the number of output targets dynamically to help select the outputs, effectively tackling the varying number of target objects in flexible referring expressions. The experimental results demonstrate the remarkable superiority and generalizability of the proposed HieA2G on multiple visual grounding tasks including REC, GREC, phrase grounding, RES, and GRES.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (62472104, 62137002, 62293550, 62293553, 62192781, and 62176209), Consulting research project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, CCF-Lenovo Blue Ocean Research Fund, Project of China Knowledge Centre for Engineering Science and Technology, Foundation of Key National Defense Science and Technology Laboratory (6142101210201), the Fundamental Research Funds for the Central Universities (xhj032021013-02).

References

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deng, J.; Yang, Z.; Liu, D.; Chen, T.; Zhou, W.; Zhang, Y.; Li, H.; and Ouyang, W. 2023. TransVG++: End-to-end visual grounding with language conditioned vision transformer. *IEEE TPAMI*.

Ding, H.; Jiang, X.; Shuai, B.; Liu, A. Q.; and Wang, G. 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2393–2402.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023a. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2694–2703.

Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P. H.; and Bai, S. 2023b. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20224–20234.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16321–16330.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2023c. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7900–7916.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, S.; and Ding, H. 2024. RefMask3D: Language-Guided Transformer for 3D Referring Segmentation. In *ACM International Conference on Multimedia*, 8316–8325.

He, S.; Ding, H.; Jiang, X.; and Wen, B. 2024. SegPoint: Segment Any Point Cloud via Large Language Model. In *European Conference on Computer Vision*, 349–367.

He, S.; Ding, H.; Liu, C.; and Jiang, X. 2023. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Kim, N.; Kim, D.; Lan, C.; Zeng, W.; and Kwak, S. 2022. ReSTR: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, L.; Bu, Y.; and Cai, Y. 2021. Bottom-up and bidirectional alignment for referring expression comprehension. In *ACM MM*, 5167–5175.

Li, M.; and Sigal, L. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34: 19652–19664.

Li, X.; Ding, H.; Yuan, H.; Zhang, W.; Pang, J.; Cheng, G.; Chen, K.; Liu, Z.; and Loy, C. C. 2024. Transformer-Based Visual Segmentation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10138–10163.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10880–10889.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, C.; Ding, H.; and Jiang, X. 2023. GRES: Generalized Referring Expression Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23592–23601.

Liu, C.; Jiang, X.; and Ding, H. 2024. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2(1): 16.

Liu, Q.; Wen, Y.; Han, J.; Xu, C.; Xu, H.; and Liang, X. 2022. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, 275–292.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019a. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 *Proceedings of the IEEE/CVF international conference on computer vision*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10034–10043.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In *CVPR*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571.

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 *Proceedings of the IEEE International Conference on Computer Vision*, 2641–2649.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. CRIS: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11686–11695.

Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; Ghanem, B.; and Tao, D. 2024a. Towards Open Vocabulary Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 5092–5113.

Wu, Z.; Weng, Z.; Peng, W.; Yang, X.; Li, A.; Davis, L. S.; and Jiang, Y. 2024b. Building an Open-Vocabulary Video CLIP Model With Better Architectures, Optimization and Data. *IEEE TPAMI*.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18134–18144.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022a. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9499–9508.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic graph attention for referring expression comprehension. In *ICCV*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022b. LAVT: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18155–18165.

Ye, J.; Lin, X.; He, L.; Li, D.; and Chen, Q. 2021. One-stage visual grounding via semantic-aware feature filter. In *ACM MM*, 1702–1711.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2024. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *International Conference on Learning Representations*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *ECCV*.

Yu, T.; Hui, T.; Yu, Z.; Liao, Y.; Yu, S.; Zhang, F.; and Liu, S. 2020. Cross-modal omni interaction modeling for phrase grounding. In *ACM MM*.

Zhou, Y.; Ji, R.; Luo, G.; Sun, X.; Su, J.; Ding, X.; Lin, C.-W.; and Tian, Q. 2021. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*.

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 *European Conference on Computer Vision*, 598–615.

Zhuang, B.; Wu, Q.; Shen, C.; Reid, I.; and Van Den Hengel, A. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4252–4261.