# Generalized DINO : Detection Transformers via Multimodal Models for Generalized Object Detection

Yuen, Ka Shing / 阮, 嘉誠

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
19
(開始ページ / Start Page)
1
(終了ページ / End Page)
7
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030602

# Generalized DINO: Detection Transformers via Multimodal Models for Generalized Object Detection

Ka Shing Yuen
Graduate School of Computer and Information Sciences
*E-mail: kashing.yuen.5j@cis.k.hosei.ac.jp*

## Abstract

*Referring Expression Comprehension (REC) is a task in the realm of vision and language, aiming to identify objects in images based on provided descriptions. Classic REC methods, however, face challenges in handling expressions involving multiple targets or empty scenarios. In this paper, we study the limitations of existing REC methods, particularly in the context of Generalized Referring Expression Segmentation (GRES). In response, we propose Generalized DINO, a model that extends Transformer-based detectors by incorporating Region-Image Cross Attention (RIA) and Region-Language Cross Attention (RLA) mechanisms. This approach enables the detector to support arbitrary numbers of target object detection, overcoming the constraints of traditional REC methods. Comprehensive experiments on widely-used datasets such as RefCOCO/+/g and the GRES benchmark gRefCOCO showcase the superior performance of Generalized DINO in GRES tasks. The model outperforms even the robust RELA model, demonstrating a significant stride in handling expressions with multiple targets or empty scenarios. Our findings underscore the efficacy of Generalized DINO in enhancing the robustness and flexibility of REC models, contributing to multimodal information processing. The model's ability to handle complex language expressions involving multiple objects positions it as a valuable asset in applications like human-computer interaction and visual question answering.*

## 1. INTRODUCTION

Object detection, a fundamental concept in computer vision, revolves around the ability to train machines to recognize and precisely locate objects within images or videos. In the context of this paper, imagine having a photograph or a frame from a video containing various objects like people, cars, or animals. Object detection involves empowering a computer system to identify and pinpoint the positions of these objects in the visual content.

This paper delves into the realm of object detection, with a particular focus on advancing capabilities through a task known as Referring Expression Comprehension (REC) [13, 14]. Unlike traditional object detection, which relies solely on visual features such as shapes and colors, REC integrates language understanding. It takes object detection a step further by comprehending language expressions that describe the specific locations of objects within an image [1, 2, 3].

However, the traditional REC methods have notable limitations that impede their effectiveness in real-world applications, particularly their struggle with multiple target objects and empty-target expressions. As shown in Figure 1(a) and (b) , we are using Grounding DINO, a transformer-based detector, which is traditional REC methods owned state-of-art performed in REC, still unable to deal with multiple targets and empty-targets.

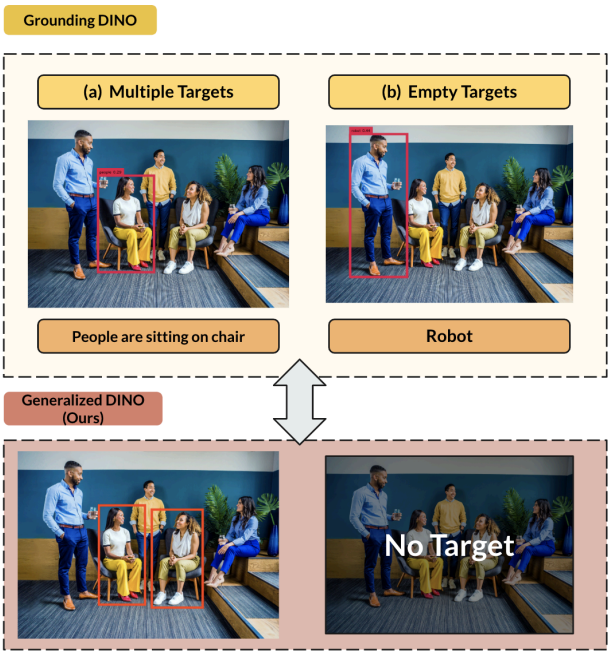


Figure 1: Comparison of the object detection result by Grounding DINO [6] and Generalized DINO in Generalized Referring Expression Segmentation (GRES) [5]. (a) Grounding DINO unable to detect all the correct targets when multiple targets are requested. However, Generalized DINO successfully bound all the target objects in the image(b) When the referent does not exist in the image, for example, the empty target is requested, Grounding DINO still detects the wrong target in the image due to the model design proposed. In contrast, Generalized DINO can reject the empty target.

Motivated by the need to bridge language comprehension with visual perception, the paper explores REC's applications in human-computer interaction, visual question answering, and video production. It introduces innovative terms and

Supervisor: Prof. Kaoru Uchida.

technologies aimed at enhancing the REC task, ultimately elevating the performance of object detection methods.

To overcome challenges faced by traditional REC methods, we propose a novel approach named Generalized DINO. This innovative method addresses limitations, such as struggles with multiple target objects and empty-target expressions. By combining techniques, such as incorporating DINO with RELA [5], aim to enhance the model's flexibility, improving its ability to accurately detect and comprehend multiple objects within a given scene. Building a network that divides the image into regions and facilitates explicit interactions between these regions. The soft-collation approach significantly improves flexibility, allowing a more nuanced treatment of features in each region. This change let Generalized DINO be able to handle multi-target detection in a single expression to make the model more robust and reliable applications in natural language processing and computer vision.

## 2. RELATED WORKS

### 2.1. Referring Expression Comprehension (REC)

Referring Expression Comprehension (REC) [13, 14] is a task that combines vision and language, where the goal is to detect specific objects in an image based on language expressions. In simpler terms, REC is about teaching machines to understand and act upon instructions in the form of spoken or written language to identify objects in pictures.

REC methods typically use either a one-stage or two-stage approach. In a one-stage approach, the system directly predicts the location of the objects in the image, while in a two-stage approach, it first generates potential regions and then predicts the final object locations. These methods, however, have limitations when dealing with expressions that point to multiple objects.

To address this limitation, a new task called Generalized Referring Expression Segmentation (GRES) [5] has been introduced. GRES aims to handle situations where users mention multiple objects in one instruction or describe things not present in the image. Unlike traditional REC, GRES is designed to address more complex scenarios involving multiple targets or cases where there are no targets in the image. It breaks away from the constraints of traditional REC methods by accommodating both scenarios of multiple targets and situations where there are no specified targets in the image, which is the main scope of our research of Generalized DINO.

### 2.2. Detection Transformers

Object detection has seen significant advancements through the use of Detection Transformers (DETRs) [7, 8, 9]. A Detection Transformer is a transformer-based model specifically which integrates attention mechanisms with modifications for object detection in computer vision. DETRs initially exhibited excellent performance in identifying objects [20], but their high computational requirements limited their practicality. Real-Time DETR (RT-DETR) addressed this challenge by introducing key enhancements to reduce computational costs. Its design incorporates a backbone, hybrid encoder, and transformer decoder, which effectively lowers the computational burden. The IoU-aware query selection further refines the model's focus on relevant objects, enabling real-time object detection on accelerated backends. The development of DETR structures, with subsequent refinements, paved the way for improvements [15,16,17].

Inspired by BYOL (Bootstrap Your Own Latent), DINO [10], a noteworthy self-supervised Vision Transformer (ViT) [21], is recognized for its ability to learn representations from unlabeled data. BYOL, is a self-supervised learning method that emphasizes learning useful representations without explicit labels. It employs a dual neural network setup where one network generates augmentations of the input data, and the other predicts these augmentations, facilitating the learning of rich representations. Building upon this foundation, DINO introduces the concept of contrastive de-noising and achieves record-breaking performance on the COCO object detection benchmark.

Grounding DINO [6], an open-set detector at multiple stages based on DINO structure with grounded pre-training, set a state-of-the-art performance in open-set object detection. However, Grounding DINO fails to tackle the challenge in GRES due to the decoder structure only fulfilling one expression that matches one existing target, which is addressed by our proposed Generalized DINO.

## 3. PROPOSED METHOD

In this section, we will introduce the model design of Generalized DINO. In the previous section, we analyze the limitations observed in Grounding DINO concerning Generalized Referring Expression Segmentation (GRES). Due to multi-target expression owning more complex attribute description and relationship, It required the model to be able to identify complex interaction among regions in the image, and capture detailed attributes for all objects.

### 3.1. Architecture Overview

Figure 2 displays the overall design of Generalized DINO, which, like Grounding DINO, uses a dual-encoder-single-decoder architecture with separate components for processing images and text [6]. It contains an image backbone for extracting features from images, a text backbone for extracting features from text, a feature enhancer for image and text feature fusion, multi-model Transformer decoder for predicted probability of target region and U-net for Box refinement.

In processing each (Image, Text) pair, the first step is extracting vanilla image features and vanilla text features, derived from image and a text backbone, are input into a feature enhancer module for cross modality feature fusion to obtain cross-modality text and image features. Next, the text and image features are sent to the multi-model decoder to predict the probability of the target region. Finally, the probability maps are fed into U-net to do the segmentation and predict object boxes. As a result, Generalized DINO outputs the object boxes from given referring expressions. For example, as shown in Figure 2, the model locates all standing people from the input image in once model calls. Meanwhile, if

the probability maps predict no target on the image after prediction, the probability maps will be set to empty and output nothing on the image.

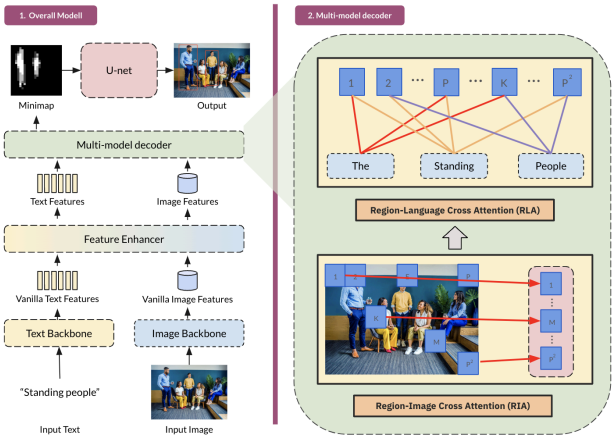


Figure. 2: The framework of Generalized DINO. We present the overall framework and multi-model decoder in block 1 and block 2, respectively.

## 3.2. Feature Extraction and Enhancer

Following Grounding DINO's idea, adding more feature fusion in the pipeline improves the model's performance [6, 11,12]. Therefore, we perform early fusion in the neck module which aims to help align features of different modalities. The input image will extract image features by Swin transformer [18] encoder, and the input language expression will extract language features by BERT [19] simultaneously. Once vanilla image and text features are extracted, we input them into a feature enhancer designed for cross-modality feature fusion including three feature enhancer layers. In the first layer, we use Deformable self-attention to enhance image features and regular self-attention to enhance text features. Additionally, we integrate image-to-text cross-attention and text-to-image cross-attention for feature fusion inspired by GLIP [20].

## 3.3. Multi-model decoder

The DINO [10] method uses a decoder to map the learned features to object detection outputs. The decoder is responsible for predicting the bounding boxes and class labels of objects in an image. However, DINO employs a query selection mechanism to generate object bounding boxes by selecting the most matching query. Since each query can only match one bounding box, DINO can only detect one result at a time. The purpose of DINO is designed for individual object detection. Therefore, classic DINO structure is unable to handle the GRES problem, which allows arbitrary numbers of referred targets, including multiple instances or no target circumstances during the detection.

However, inspired by RELA [5], Generalized DINO extends to be able to deal with multiple targets and empty targets. Generalized DINO's decoder divides the feature maps into $P \times P = P^2$ regions, and interacts among these regions in the model. Think of these "regions" as corresponding to patches in the image, much like ViT does. Each region will individually calculate the probability of expression matching and finally form a "minimap" for further use.The multi-model decoder has two main parts: Region-Image Cross Attention (RIA) and Region-Language Cross Attention (RLA).

## 3.3. Region-Image Cross Attention (RIA)

The RIA model achieves this by performing attention interaction between image features and learnable region-based queries, resulting in the generation of region features corresponding to each query. Firstly, attention maps are generated for each query by interacting the image features with query embeddings, indicating the spatial regions in the image that correspond to each query. Then, these attention maps are utilized to extract region features from the respective regions for further use.

## 3.5. Region-Language Cross Attention (RLA)

The RLA models dependencies between regions and language. It consists of self-attention and multimodal cross-attention. Self-attention is used to model the dependencies between regions by interacting a region feature with all other regions, calculating an attention matrix, and outputting relation-aware region features. Multimodal cross-attention is used to model the dependencies between regions and language. It takes language features as input for values and keys and region image features as input for queries. It establishes relationships between each word and each region and generates language-aware region features using the obtained word-region attention. Finally, the relation-aware region features, language-aware region features, and region image features are combined and further fused using a multi-layer perceptron (MLP).

## 3.6. Outputs and Loss

The output from the multi-model decoder, called a "minimap", shows the probability of each region containing targets. The minimap can be regarded as the predicted segmentation mask used in object detection. It will be input into the U-Net to perform segmentation to identify target objects at a pixel level; post-processing steps can be applied to create bounding boxes. By analyzing the segmented regions, algorithms can determine the minimum bounding rectangles or use contour information to encapsulate the identified objects. This involves finding the coordinates of segmented regions and generating bounding boxes accordingly.

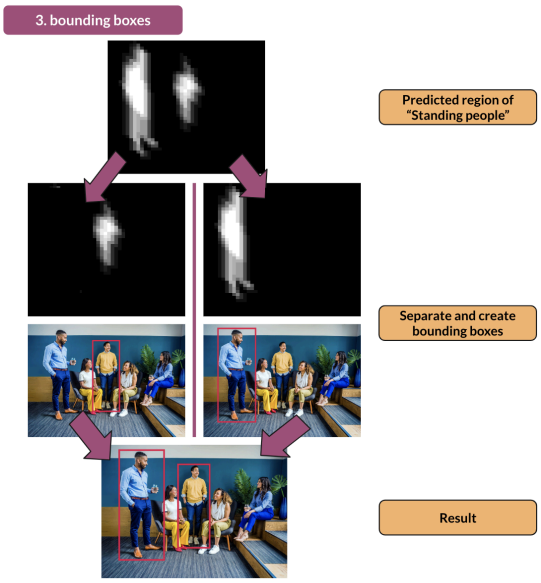


Figure. 3: U-Net performs segmentation to separate the minimap, then identify target objects and create bounding boxes to output results.

## 4. Experiments

We conduct extensive experiments to represent the efficacy of Generalized DINO. First, we show the results on the gRefCOCO dataset to show the performance of Generalized DINO in GRES tasks. Then we test Generalized DINO against other REC methods and thoroughly assess their performance.

### 4.1. Generalized Referring Expression Segmentation (GRES)

**Setting.** For conducting GRES tasks with Generalized DINO, we utilize the gRefCOCO dataset [5] for experimentation. The gRefCOCO dataset contains 278,232 expressions, including 80,022 multi-target expressions and 32,202 no-target expressions, referring to 60,287 distinct instances found in 19,994 images. All bounding boxes for the target instances are provided. Following the UNC partition of RefCOCO, we divide the images into four subsets: training, validation, test-A, and test-B. We add the gRefCOCO training set into the mixed training dataset in Generalized DINO to pretrain for 150,000 steps. Throughout this experiment, we use Generalized Intersection over Union (gIoU) and Complete IoU (CIoU) as the primary metrics for object detection, assessing the accuracy of bounding box predictions. GIoU averages the IoU for each mask, whereas cIoU computes the cumulative intersection area over the cumulative union area across the whole dataset. For no-target samples, we introduce the No-target-accuracy (N-acc.) metric to measure the model's performance on identifying no-target samples. For no-target samples, the gIoU values of true positive no-target samples are regarded as 1, while gIoU values of false negative samples are treated as 0.

**Results.** Table 1 shows the performance of different models in GRES. We mark the best results in bold. RELA is the strongest model on the GRES. But Generalized DINO outperformed RELA, showing approximately a 2% improvement in gIoU and 5% in N-acc. Generalized DINO achieved over 66% in gIoU on the validation set, showing the superiority of Generalized DINO in GRES task.

Table1 Generalized referring expression segmentation (GRES) results on gRefCOCO. gIoU and CIoU are metrics focusing on bounding box accuracy in object detection and N-acc evaluates a model's ability to correctly identify samples with no target present

| Method | Validation Set | | |
|---|---|---|---|
| | gIoU | cIoU | N-acc. |
| LTS [23] | 53.16 | 51.54 | - |
| VLT [24] | 51.73 | 51.62 | 47.47 |
| CRIS [25] | 57.19 | 55.51 | - |
| LAVT [22] | 58.14 | 57.50 | 49.72 |
| ReLA [5] | 63.06 | 62.18 | 56.10 |
| **Generalized DINO (Ours)** | **66.77** | **63.78** | **62.97** |

### 4.2. Referring Expression Comprehension (REC)

**Setting.** To evaluate its capabilities across diverse tasks, we examine the performance of Generalized DINO in the REC task. We use Grounding DINO [10] as our baseline to compare the performance of different models on RefCOCO, RefCOCO+ [6], and RefCOCOg [7]. The models are firstly pre-trained as in GRES, and then fine tuned for 10 epochs with a joint dataset of these three REC training sets. The primary metrics for object detection in this task are gIoU and CIoU.

Table2 Referring Expression Comprehension results on RefCOCO dataset

| Method | RefCOCO | | |
|---|---|---|---|
| | Val. | Test-A | Test-B |
| VGTR [26] | 77.15 | 82.09 | 69.05 |
| TransVG [27] | 79.53 | 82.95 | 73.83 |
| RefTR [28] | 80.04 | 82.77 | 77.73 |
| MDETR [29] | 86.91 | 87.84 | 81.70 |
| Grounding DINO [6] | **90.12** | **92.77** | **87.28** |
| **Generalized DINO (Ours)** | 87.93 | 90.51 | 83.99 |

**Results.** Table 2 shows the performance of different models in the REC task. We mark the best results in bold. Our model demonstrates nearly identical performance to Grounding DINO, which is the second best accuracy performance on the REC task. However, in the test set of RefCOCO, Grounding DINO outperforms our model by approximately 2% to 5%. Further discussion on this performance difference is provided in the discussion.

### 4.3. Visualization

Figure 3 displays visual results of Generalized DINO from the validation set of gRefCOCO, illustrating how the model addresses the primary challenges in GRES—multiple targets and empty targets. In part (a), Generalized DINO successfully bound the box on all targets referred to, while Grounding DINO [6] only bound one of requested instances. In part (b), Grounding DINO mistakenly bound the box on the object in

the image which disagrees with the referring expression, but Generalized DINO successfully rejects all the empty targets.


Figure. 4: Visualizations of Grounding DINO [6] and Generalized DINO in the GRES task [5]. The first column shows the referring expressions in the instructions, the second column shows Grounding DINO detection results, and the third column is the detection result and rejections of Generalized DINO. In the upper part is multiple targets cases, each target will bound a box on the target. In the lower part is an empty target case, the images will turn dark to indicate that no box is bound in image. All examples are selected from the gRefCOCO validation set.

### 4.4. Discussion and Limitations

Generalized DINO exhibits considerable strengths in handling GRES tasks; it surpasses the original best model in GRES by achieving approximately a 2-5% improvement in accuracy for GERS tasks. However, there are acknowledged limitations in traditional REC tasks, with potential challenges related to fine-grained details, complexity, and generalization. Our model, utilizing U-Net for minimap segmentation to identify target objects and bounding boxes, faces a drawback where this process may result in the loss of fine-grained details, consequently impacting box prediction accuracy.

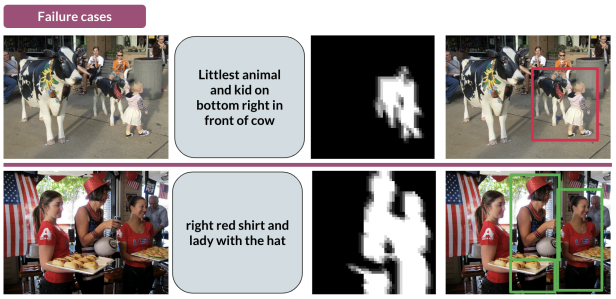


Figure. 5: Failure cases of our method on gRefCOCO dataset. Showing that the limitation of segmentation on the model

In Figure 5, we present instances where our model fails to identify target objects accurately through segmentation in the minimap. This limitation affects the overall performance of Generalized DINO on REC tasks compared to Grounding DINO, a model known for its state-of-the-art performance in object detection, primarily due to the segmentation steps involved. To address this challenge in future work, one potential approach is to minimize the impact of the segmentation process. This can be achieved by exploring alternatives such as skipping the segmentation process in the model and using a Transformer to replace U-Net, thus mitigating the loss of fine-grained details in target objects.

### 4.5. Future Work

To enhance the model's performance in GRES tasks, further work should focus on developing strategies to overcome the existing limitations. For instance, exploring advanced multimodal large language models to replace GLIP could contribute to improved results. Additionally, addressing cases where users reference multiple subjects in a singular prompt or provide incongruent descriptions with any image target remains a challenge. Future research efforts should aim to refine the model's ability to handle such complex scenarios, potentially involving more sophisticated attention mechanisms or context-aware approaches.

## 5. Conclusion

We have presented a Generalized DINO in this paper, specifically designed to overcome limitations inherent in existing Referring Expression Comprehension (REC) methods, particularly in the challenging context of Generalized Referring Expression Segmentation (GRES) [5]. By extending the Transformer-based detector DINO with the Region-Image Cross Attention (RIA) and Regional Language Attention (RLA) mechanism, Generalized DINO demonstrates remarkable improvement in handling expressions involving multiple targets or empty scenarios. Our experiments conducted on gRefCOCO datasets, highlight the superior performance of Generalized DINO in GRES tasks, surpassing even the robust RELA model. Generalized DINO represents a significant step towards enhancing the robustness and flexibility of REC models, contributing to the broader field of multimodal information processing.

## REFERENCES


[1] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In IEEE CVPR, 2021.

[2] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In IEEE CVPR, 2020

[3] Qie Sima, Sinan Tan, Huaping Liu, Fuchun Sun, Weifeng Xu, and Ling Fu. Embodied referring expression for manipulation question answering in interactive environment. In IEEE ICRA, 2023.

[4] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In EMNLP, 2014.

[5] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In IEEE CVPR, 2023.

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023.

[7] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13619–13627, 2022.

[8] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. 2023.

[9] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In International Conference on Learning Representations, 2022.

[10] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.

[11] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. computer vision and pattern recognition, 2017.

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. arXiv preprint arXiv:2112.03857, 2021.

[13] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. international conference on computer vision, 2017.

[14] Peihan Miao, Wei Su, Lian Wang, Yongjian Fu, and Xi Li. Referring expression comprehension via cross-level multimodal fusion. ArXiv, abs/2204.09957, 2022.

[15] Qiang Chen, Xiaokang Chen, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. 2022. 3

[16] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7373–7382, 2021.

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544, 2021.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021.

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

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. arXiv preprint arXiv:2112.03857, 2021.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In Proc. Int. Conf. Learn. Represent., 2021.

[22] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In IEEE CVPR, 2022.

[23] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate the segment: A strong pipeline for referring image segmentation. In IEEE CVPR, 2021.

[24] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In IEEE ICCV, 2021.

[25] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In IEEE CVPR, 2022.

[26] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. 2021.

[27] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. arXiv: Computer Vision and Pattern Recognition, 2021.

[28] Muchen Li and Leonid Sigal. Referring transformer: A onestep approach to multi-task visual grounding. arXiv: Computer Vision and Pattern Recognition, 2021.

[29] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel

Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrmodulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1780–1790, 2021.