INSTRUCTDET: DIVERSIFYING REFERRING OBJECT DETECTION WITH GENERALIZED INSTRUCTIONS

Ronghao Dang[1*] **Jiangyan Feng**[2*] **Haodong Zhang**[2] **Chongjian GE**[3] **Lin Song**[4]
Lijun Gong[2] **Chengju Liu**[1] **Qijun Chen**[1] **Feng Zhu**[2] **Rui Zhao**[2] **Yibing Song**[5⋆]
[1]Tongji University [2]SenseTime Research [3]The University of Hong Kong
[4]Tencent AI Lab [5]Alibaba DAMO Academy
dangronghao@tongji.edu.cn fengjiangyan@sensetime.com yibingsong.cv@gmail.com

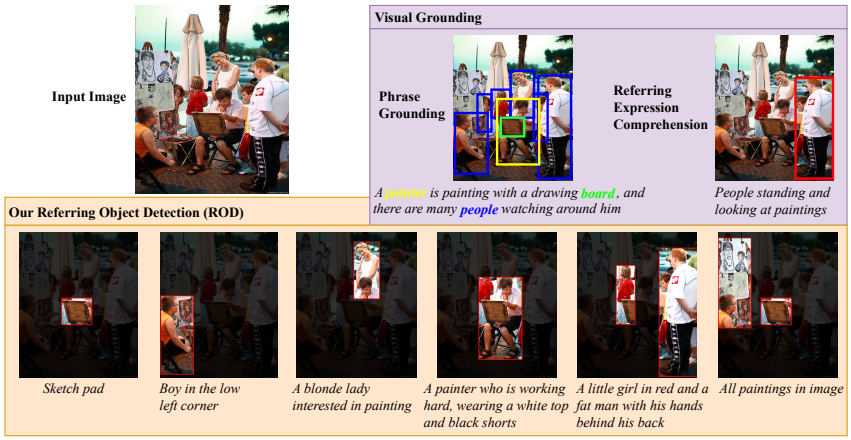


Figure 1: Our ROD aims to execute diversified user detection instructions compared to visual grounding. For images with object bbxs, we use foundation models to produce human-like object detection instructions. By training a conventional ROD model with incorporating tremendous instructions, we largely push ROD towards practical usage from a data-centric perspective.

ABSTRACT

We propose InstructDET, a data-centric method for referring object detection (ROD) that localizes target objects based on user instructions. While deriving from referring expressions (REC), the instructions we leverage are greatly diversified to encompass common user intentions related to object detection. For one image, we produce tremendous instructions that refer to every single object and different combinations of multiple objects. Each instruction and its corresponding object bounding boxes (bbxs) constitute one training data pair. In order to encompass common detection expressions, we involve emerging vision-language model (VLM) and large language model (LLM) to generate instructions guided by text prompts and object bbxs, as the generalizations of foundation models are effective to produce human-like expressions (e.g., describing object property, category, and relationship). We name our constructed dataset as InDET. It contains images, bbxs and generalized instructions that are from foundation models. Our InDET is developed from existing REC datasets and object detection datasets, with the expanding potential that any image with object bbxs can be incorporated through our InstructDET method. By using our InDET dataset, we show that a conventional ROD model surpasses existing methods on both standard REC datasets and our InDET test set. InstructDET, our data-centric method with automatic data expansion by leveraging foundation models, directs a promising field that ROD can be greatly diversified to execute common object detection instructions.

[*]R. Dang and J. Feng contribute equally. [⋆]Y. Song is the corresponding author. This work is done when R. Dang is an intern at Sensetime. The code is available at https://github.com/jyFengGoGo/InstructDet.

1 INTRODUCTION

Referring object detection (ROD) aims to detect target objects according to language reference that represents user intentions. ROD is closely related to visual grounding where there are phrase grounding (Akbari et al., 2019; Li et al., 2022a; Gao et al., 2023) and referring expression comprehension (Su et al., 2020; Zhu et al., 2022). As shown in Fig. 1, phrase grounding detects all objects mentioned in one sentence, while referring expression comprehension (REC) only detects one single object that the text refers to. As such, the language reference in REC shall be discriminative and specifically relates to one object without ambiguity.

Currently, visual grounding develops at an initial stage and leaves a gap for practical usage. The phrase grounding does not differentiate which object ought to be detected via language description, while REC only targets for one object with single text reference. In the current REC datasets, each image contains few expressions (e.g., 1 or 2 phrases). These expressions are insufficient to represent user intentions. In an image where there are several objects, users may want to detect each single object by using different descriptions (e.g., object color, shape, or location), or detect multiple objects in different combinations (e.g., similar properties or relationships). These diverse expressions are not conveyed within current REC datasets, leaving the gap for existing methods to practically fulfill user intentions for visual grounding. Moreover, the manual collection of these expressions are cumbersome, and subject bias prevents an effective coverage of common user intentions when perceiving each image. Therefore, the practical user expressions are not well fulfilled when they expect to detect various objects in one image.

In this work[1], we aim to push visual grounding toward practical usage from a data-centric perspective. Instead of developing REC models to generalize based on current data, we set up referring object detection (ROD) scenario to automatically diversify user expressions. Inspired by the generalization of foundation models that execute common user instructions based on the image and text inputs, our InstructDET borrows their capabilities to produce human-like instructions that encompass user intentions related to object detection. The generalized instructions produced by the foundation models can be regarded as an expansion of existing user expressions in REC. We produce instructions that describe single object from two pipelines. In the first pipeline (i.e., global prompt), we convert an image into an elaborate text description via LLaVA (Liu et al., 2023a). The text description, together with object bbxs coordinates, are sent to the LLaMA (Touvron et al., 2023) for instruction generation in global prompt. During generation, we manually write 3 in-context examples and leverage the in-context learning (Dong et al., 2023) ability of LLaMA to describe the content related to each object following the format of our examples.

In the second pipeline (i.e., local prompt), we send the image and text prompts into LLaVA. The objects in the image are marked with bbxs and the text prompts require LLaVA to describe the object content. We initialize LLaVA with miniGPT4 weights and find it tends to produce lengthy and global descriptions. So we perform a partial finetuning on LLaVA by using REC data to let it focus on local objects. Through these two pipelines, we observe that instructions generated from global prompt pipeline focus more on the object relationship, while instructions generated from local prompt pipeline focus more on rich visual details and advanced logic reasoning. Naturally, we combine instructions from these two pipelines to formulate expressions for single referred object. During instruction generation, the uncontrolled model hallucination (Li et al., 2022b) brings incorrect or irrelevant instructions. We propose to use visual-textual verification via CLIP (Radford et al., 2021) for effective instruction filtering.

The generalization and reasoning of foundation models (Wang et al., 2022; Zhou et al., 2022) provide sufficient instructions encompassing user intentions for single object description. When describing multiple objects, we divide descriptions into two parts. The first part is to independently describe each single object followed by concatenation, and the second part is to summarize commonalities of multiple objects. The commonality summarization requires unifying similar or related objectives by a higher-level language abstraction that describes their similarities and relationships. We collect the

[1]We do not differentiate "instruction" and "expression" in this paper, as both of them represent user intentions. For presentation clarity, in our InstructDET pipeline we refer expressions that are generated by foundation models, and we further refine expressions to instructions for InDET inclusion. As we only focus on ROD, we can formalize our instruction by simply adding the word 'detect' beforehand.

combinations of different objects via semantic clustering, then utilize LLM to generate commonality summarizations for each combination.

We automatically collect instructions targeting for single or multiple objects in images and construct our InDET dataset. Sec. 4 shows an in-depth analysis of our dataset where we establish a guideline to organize these instructions from 6 aspects. Compared to existing REC datasets where the instructions only reside in sub-parts of our groups, our InDET is more comprehensive to incorporate user intentions of object detection. Fig. 1 shows an intuitive example of the generalized expressions produced by foundation models. By using our InDET dataset, we train a conventional ROD model and find it surpasses existing VG models on standard benchmarks and our InDET test set. Moreover, we also validate that our model has learned to effectively understand the meaning of instructions rather than only recognize key words, which is because of the tremendously expressive instructions incorporated for our model training. Our InstructDET method can automatically expand training data by using in-the-wild images with object bbxs, which improves our model generalizations towards practical usage. In addition, our model can already serve as the detection module of the neural-symbolic visual compositional task solution given arbitrary language instructions beyond object detection (e.g., Visual ChatGPT (Wu et al., 2023), VISPROG (Gupta & Kembhavi, 2023)).

2 Related Works

Visual Grounding. Studies on visual grounding (Kamath et al., 2021; Chen et al., 2021; Deng et al., 2021; Su et al., 2023) can be mainly categorized as phrase grounding (Plummer et al., 2022; Kojima et al., 2023) and REC (Hudson & Manning, 2018; Li & Sigal, 2021). Phrase grounding detects all objects mentioned in the text while REC localizes one object that the text referred to. In (Zhang et al., 2022; Liu et al., 2023c), the objects mentioned in the text are verified to each visual object proposal one-by-one. These methods require a clear and specific object referring in the text. On the other hand, methods (Zhu et al., 2022; Yan et al., 2023) based on DETR (Carion et al., 2020) can accept abstract and summarized descriptions such as "red objects" and "all objects". Our ROD model follows DETR-based design to enrich interpretation of various instructions. Note that our model is learned via InDET dataset where instructions are produced based on preset object bbxs.

Referring Expression Datasets. The REC datasets are usually constructed via manual annotation on the images. A two-player game is utilized in (Kazemzadeh et al., 2014) where the text descriptions are concise due to limited relevant visual contents. The RefCOCO, RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) employ MSCOCO (Lin et al., 2014) images for manual expression production. The expression flexibility and diversity of these datasets are limited to encompass common detection intentions. Recent datasets (Krishna et al., 2017; Kuznetsova et al., 2020; Kebe et al., 2021) focuses on data scalability rather than auto generation. Cops-Ref (Chen et al., 2020) leverages scene graph as reasoning groundwork, thus forming a tree structure to generate expressions with varying compositionality. Different from these methods based on template guided expression generation, our InstructDET relies on foundation models to produce well generalized and human-like instructions.

Data Generation via Foundation Models. The InstructGPT (Ouyang et al., 2022) and GPT4 (OpenAI, 2023) have shown generalization and reasoning abilities for data generation. LLaVA (Liu et al., 2023a) first uses GPT4 for multi-modal data generation following instructions. Otter (Li et al., 2023a) performs multi-modality in-context instruction tuning by levering multiple images, questions, and answers. Currently, these models focus on global image and language understanding, with less focus on local object analysis. Moreover, these multi-modality models, although processing multi-modality data, still outputs single-modality text description. There is a gap for these foundation models to function in the computer vision scenarios, especially visual recognition. In comparison, our InstructDET uses foundation models to benefit ROD model training, which contributes directly to improve object detection performance.

3 InstructDET

Fig. 2 shows an overview of our InstructDET method for data construction. Given an input image with object bbxs, we use two pipelines to produce detection expressions from foundation models. The expressions are further refined to instructions and incorporated into our InDET dataset.

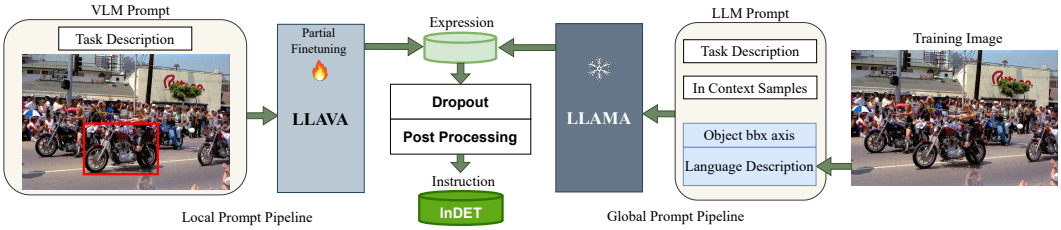


Figure 2: An overview of our InstructDET. We use two pipelines to produce detection expressions via foundation models. In the global prompt pipeline, we use LLaVA to describe an image via text, and combine this text with other text prompts for LLaMA input. In the local prompt pipeline, we use the same image with object bbxs and text prompts as multi modality input for LLaVA. The produced expressions are further refined to instructions and incorporated into our InDET dataset.

3.1 Global Prompt Pipeline

The large language model (LLM) has shown surprising generalizations to well execute common user instructions. We use LLM to simulate user intentions when perceiving objects in an image. Our global prompt pipeline produces a text prompt for the LLaMA[2] model. This prompt consists of several contents including global image description, object bbx coordinates, in-context samples, and task description. Without tuning LLaMA, we obtain instructions that describe objects in an image. A detailed example is shown in Sec. C for an intuitive illustration of how we leverage foundation models to produce expressions. We elucidate the key steps during this process as follows:

Given an image with object bbxs, we first obtain global image description in text form. If this image already contains dense captions (e.g., from Flicker30K), we directly load these captions. Alternatively, we leverage LLaVA to generate the global image description. The text prompt we use for LLaVA contains our language guidance to emphasize that specific interested object categories shall be mentioned. As such, LLaVA will describe each labeled object in its output. As for object bbx content, if the image is from REC dataset, we use referring expression as the object content. Otherwise, we simply use the category name.

When designing the task description prompt, we expect LLaMA to produce diverse expressions that contain different properties of single object as much as possible. We manually list the attributes from the user perspective, including the object type, color, function, motions, etc. Besides, we include the object attributes of its relationship with other objects in image, such as object interactions, object relative positions, etc. When using these aforementioned prompts for LLaMA input, we find that the output text varies significantly and might be irrelevant to the target objects. Inspired by the in-context learning ability of foundation models, we manually design in-context samples to regularize the output content and format. The output results will thus resemble our in-context examples but with our expected diversified object descriptions.

3.2 Local Prompt Pipeline

The global prompt pipeline produces expressions according to text prompts. Naturally, we can feed both image and text prompt to the multi-modality foundation model for object description. Given an image, we mark the object with bbx rectangle, and send this image to LLaVA, together with the text prompt that requires LLaVA to describe the object according to the bbx. Here, the bbx serves as a visual highlight for LLaVA to comprehend the target object that we expect to describe.

Our LLaVA model is initialized with miniGPT4 weights. When we send these multi-modality inputs to LLaVA, we observe that LLaVA produces detailed and dense descriptions for the image, rather than expressions of the specific target object. We analyze that the vision-language alignment module in LLaVA is the Q-Former (Li et al., 2023b), which transforms one image into only 32 visual tokens without concentrating on the local objects. Meanwhile, LLaVA itself tends to produce lengthy and dense descriptions. In order to generate instructions suitable for ROD, we finetune a part of LLaVA by using existing REC datasets. Specifically, we only update a linear layer that transforms visual

[2]In this paper, we use a variant of LLaMA (i.e., Vicuna 13B) that has gone through instruction tuning. Besides, we use a variant of LLaVA, which is a multi-modal paradigm that maps visual features into token embeddings with further alignment to the text domain.

tokens to the text embedding space during training. The linear layer is learned to attend local objects with concise expressions. After finetuning, we observe the LLaVA output becomes informative and closely related to the target object. Detailed examples of generating expressions in local prompt pipeline are shown in Sec. B, and the detailed analysis on how finetuning improves LLaVA output is provided in Sec. H.

3.3 Expression Filter

In global and local pipelines, we have regularized the output of foundation model from several aspects including text prompt specification, in-context learning, and model finetuning. In practice, we still observe the model hallucination phenomena that the model sometimes generate expressions describing objects not even exist in the image. Moreover, the expressions from the local prompt pipeline sometimes describe the whole image rather than local objects. This is due to the miniGPT4 initialization of LLaVA, which utilizes dense captions for instruction tuning. The tendency to generate global image description is mitigated via our model finetuning to focus on local object, but not completely disappeared. To further improve the expression quality, we introduce visual and language matching via CLIP (Radford et al., 2021) to filter out inappropriate expressions. Fig. 3 shows an overview. It contains image visual prompting and visual-textual matching.

Visual Prompting We study visual language pretraining (VLP) (Yang et al., 2023; Shtedritski et al., 2023) where visual prompting is developed for images. We observe that in zero-shot REC, coupling VLP with visual prompts enables robust pairing of local image region and corresponding text description. In the pairing process, the design of visual prompting heavily influences the visual-textual matching results. Specifically, we employ the superposition of a red ellipse and the target Gaussian blur reversion as visual prompts. A detailed pipeline illustrating visual prompting is in Sec. D.

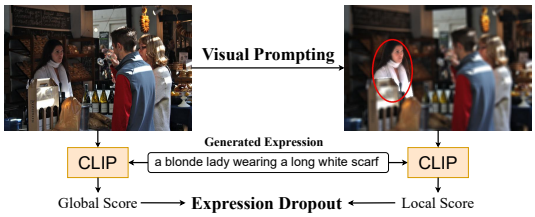


Figure 3: Expression filtering by image visual prompting and visual-textual matching via CLIP.

Visual-Textual Matching. We use images with visual prompting that emphasizes target objects to verify the corresponding text descriptions via a frozen CLIP model. While local object contents are well aligned with target referring expressions, we observe that expressions describing the whole image are not eliminated by CLIP. We analyze that CLIP is originally trained to focus on the correlations between global image features and global textual semantics. The global visual-textual matching makes CLIP model to prefer global image description accordingly. To remove this effect, we establish a referring measurement from both local and global perspectives. For the image in Fig. 3, we compute a global score S_g and a local prompt score S_l. The magnitude of referring expression can be measured via our local enhancement score $S_e = S_l - S_g$. Our final expression evaluation score can be computed as:

$$S_f = \alpha_1 S_e + \alpha_2 S_l = (\alpha_1 + \alpha_2) S_l - \alpha_1 S_g = S_l - \alpha_1 S_g \tag{1}$$

where α_1 and α_2 are scalars balancing the contributions of S_g and S_l with $\alpha_1 + \alpha_2 = 1$. So $\alpha_1 \in [0, 1]$ adjusts the final score towards local content referring or global semantics. Note that we introduce S_e to measure difference between local and global scores. If the expression is more related to the target object, S_e becomes higher after visual prompting for object highlights. After computing S_f, we set a dynamic threshold to filter out expressions. This is because S_f is based on CLIP's preference that a small target object with well matched expression achieves a lower score than a large object with mismatched expression. Therefore, we use provided expression (for images from REC) or category name (for images out of REC) to compute a final score, and discard generated instructions whose S_f is lower than this score.

3.4 Multi-Objects Expression Generation

Our expression generation pipeline illustrated above targets for each object independently. In practice, users may refer to multiple objects in one image. We study common user expressions for multi-objects, and conclude them into two aspects. The first one contains splicing expressions that

combine different single object expressions with 'and' or comma. In this case, the objects mentioned in the expression are not related to each other. The second aspect contains generalization expressions that summarize the common properties of multiple objects (e.g., color, category, or location) to produce an abstract and conceptual description. It resembles mining similarities between multiple objects and thus is not straightforward to conclude. Therefore, we need to discover object combinations where similar properties may exist, and then summarize the commonalities among them to constitute the summarized expressions.

Our process to produce summarized expression is shown in Fig. 4. For each object, we first concatenate all its related expressions with commas. Through this concatenation, we can obtain this object expression from different perspectives (i.e., different properties). Then, we use the text encoder in BERT (Devlin et al., 2018) to map this concatenated expression to a semantic embedding space. As such, we obtain embeddings of concatenated expressions from all objects. Then, we cluster these embeddings into indeterminate number of clusters by using DBSCAN (Ester et al., 1996) method. We use LLaMA to generate text for clusters with multiple objects. The details of using LLaMA to mine object commonalities are in Sec. E. The generated text indicates the summarized expression we aim to produce for multiple objects.

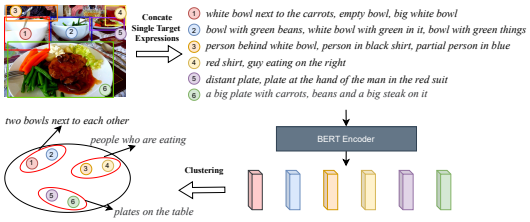


Figure 4: Mining commonalities among multi-objects via expression concatenation and text semantic clustering, followed by LLaMA descriptions on each cluster center.

Post Processing. After generating expressions for single and multiple objects, we verify and remove the repeated expressions that pertain to the same object. Then, we utilize LLaMA to further diversify these generated expressions while preserving their original intents, i.e., we use LLaMA to do synonymous rewriting of generated expressions. The prompt we use for synonymous rewriting is provided in Sec. I. We observe that for different objects in one image, the expression for one object may be similar to that of others. These expressions are ambiguous since we can not refer to a unique object based on their referring. Nevertheless, we transfer these expressions to refer to multi-objects since they express a set of objects in one image. This transfer further augments multi-object referring expressions. Finally, we collect these remaining expressions after post processing as instructions. Together with corresponding object bbxs and images, we construct our InDET dataset by incorporating diversified object detection instructions encompassing user intentions.

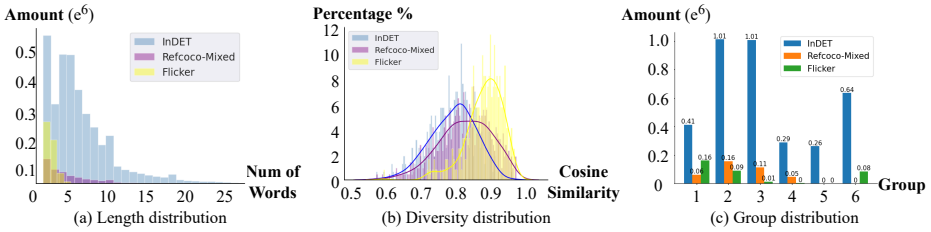
4 Dataset Analysis

Our InDET dataset contains images from MSCOCO (Lin et al., 2014), Flicker (Plummer et al., 2015), and Objects365 (Shao et al., 2019). There are 120.6K images with 908.4K referring object sets in total. Together with original expressions, there are 3.6M instructions in total, making InDET the largest real-world REC dataset at present. The average instruction length is 6.2 words and the vocabulary size is 63k words, which surpasses existing automatically annotated datasets in terms of instruction quantity, richness, and vocabulary breadth. We split the images into training, validation, and testing sets, with the corresponding instruction amount of 3139K, 240K, and 247K, respectively. In the following, we first propose a guideline that represent common user intentions and divides existing instructions into 6 groups. Then, we analyze all the instructions in InDET according to this guideline to show how our InDET advances REC scenario compared to existing datasets.

Instruction Guideline. The instructions in InDET dataset describe objects from various perspectives. We observe that these descriptions all focus on object category, attribute, and relations, but with different emphasis extent. Based on expression complexity, we establish a guideline that divides all instructions into 6 groups. Each group reflects one level of emphasis on category, attribute, and relations. Table 1 shows our guideline and examples. The first four groups are for single object and the last two groups are for multiple objects. In the first group (G1), there is only one single phrase to describe object category, which is similar to the traditional object detection task. From G2 to G4, more phrases are involved to describe the target object. For G5, we construct a spliced form to combine instructions from different single objects. In G6, the instruction is a general description

Table 1: Instruction guideline and samples. Our guideline contains 6 aspects and covers common user intentions. These aspects are built upon object category, attribute, and relations with different emphasis levels. We use ⋆ and ⋆⋆ to indicate the description complexity of different aspects.

Aspect		Category	Attribute	Relation	Examples
Single Object	1	⋆			*pencil; two children; soccer ball; city street*
	2		⋆		*shirts with English letters; red and white airplane*
	3		⋆⋆ ⋆	⋆ ⋆⋆	*man in blue shirt halfway on screen;* *people who are sitting under an umbrella;* *a man in a grey sweater and black jeans performing a skateboarding trick;*
	4		⋆⋆ ⋆ ⋆⋆	⋆ ⋆⋆ ⋆⋆ ⋆⋆	*a woman sitting cross-legged on the couch with her back facing the viewer. She has a white shirt and black pant;* *the glasses are sitting on top of some kind of paper or folder and there is a book and a lantern next to it*
Multiple Objects	5	Single object combination			*a black hat on a man's head and red umbrella and blue truck in rains*
	6	Commonality generalization			*every object on table; kids playing with the blond boy*



(a) Length distribution (b) Diversity distribution (c) Group distribution

Figure 5: Dataset analysis of expression length, diversity and group distributions.

of commonality between multiple objects. To this end, the instructions from G1 to G6 gradually introduces semantic understanding, visual language grounding, and logic reasoning for ROD. After guideline construction, we use LLaMA to assign each instruction into our groups by using in-context learning that let LLaMA to understand assigning principles and in-context assigning examples. The detailed usage of LLaMA for instruction assign is shown in Sec. F.

Instruction Length, Diversity, and Aspect Ratio Distributions. We analyze our InDET from the instruction length, diversity, and ratio distribution in our guideline groups. The RefCOCO and Flicker datasets are introduced for comparison. Fig. 5(a) shows the number of word distribution where the instruction of InDET contains more words than the other two datasets. Moreover, there are 100K instructions in our InDET consist of more than 10 words, while other datasets do not contain such informative expressions. In Fig. 5(b), we show diversity comparison where we use CLIP to map all instructions into a semantic space. Then, for the same target objects we compute average pairwise cosine similarity. The results show that our InDET contains lower value than other datasets, which indicates that our instructions are more diverse when describing the same target object. In Fig. 5(c), we show aspect ratio distribution of expressions assigned by our guideline. For existing datasets, user expressions commonly exist in G1 and G2. In contrast to the user expressions that seldom exist from G3 to G5 for Flicker, and seldom exist in G5 and G6 for RefCOCO, the instructions in our InDET exist normally in all groups. This distribution shows that our InDET is more effective to encompass common user intentions, especially for multiple objects. By leveraging our InDET, the ROD model becomes more practically applicable.

5 Referring Object Detection

In this section, we illustrate our model design for ROD task. We notice that ROD shares little difference with visual grounding (VG). First, ROD produces uncertain number of object bbxs (i.e., 0, 1, or multiple) based on one input instruction, as shown in Fig. 1. Second, ROD supports abstract and summarized object descriptions (e.g., "all objects on the table") that do not clearly refer to specific objects such as "bottle", "orange", and "knife". As recent VG models (Zhang et al., 2022; Liu et al., 2023c) require a one-by-one verification between visual objects and expression words, they are not able to execute such instructions. Motivated by the difference, we set up a conventional framework from DETR-based VG models (Zhu et al., 2022; Yan et al., 2023). Fig. 6 shows an overview of our DROD model. We illustrate key steps as follows:

Given an image with text instruction, we use visual and text encoders (Dosovitskiy et al., 2021; Devlin et al., 2018; Ge et al., 2023) to obtain their embeddings. Then, we use a bi-directional cross

attention module to perform multi-modality embedding fusion. For the fused visual embedding, we sent it to the transformer encoder and decoder structure (Zhu et al., 2020) with N learnable queries as position priors (Meinhardt et al., 2022; Ge et al., 2021). Then, the decoder produces N instance proposals for further selection. For the fused text embedding, we pass it through a global average pooling and MLP for text2visual embedding space mapping. Finally, we use cosine similarity to match proposals and mapped text embedding. During the training stage, we use confidence loss and localization loss via supervised learning. During the inference stage, we select proposals whose matching scores are above a predefined threshold, which allows our model to produce arbitrary number of bbxs for diversified instruction execution. More details are shown in Sec. G.

6 EXPERIMENTS

We evaluate the ROD performance on standard VG benchmarks (i.e., RefCOCO, RefCOCO+, and RefCOCOg) and our InDET dataset. As illustrated in Sec. 4, the images with marked objects of our InDET dataset are collected from existing datasets while the instructions are significantly enriched. We split the training and test set of InDET following RefCOCO/g/+ where the test set contains 6.5k images with an increased number of instructions to 315K. Moreover, these instructions are assigned to 6 groups according to our guideline. The performance on each group reflects how VG methods perform when processing different aspects of user instructions. The comparing methods in our experiments are from recent VG methods including MDETR (Kamath et al., 2021), Grounding-DINO (Liu et al., 2023c) and UNINEXT (Yan et al., 2023). Due to page limit, we show the evaluation results on our InDET, our InDET with shuffled expression, and standard benchmarks. Model training, and ablation studies on partial LLaVA finetuning and visual prompt selection are provided in Sec. G and Sec. H. We also provide visual comparison results of these methods in Sec. A. A video demo showing the practical usage of our DROD model is in our webpage.

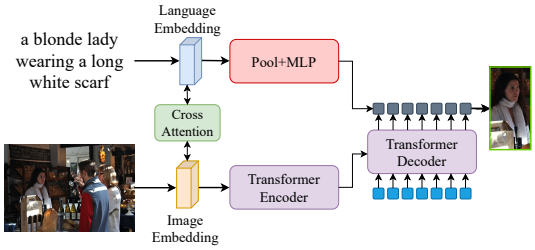


Figure 6: An overview of our diversified referring object detection (DROD) model.

Table 2: Evaluation results on our InDET and shuffled InDET test sets. We show the object bbx average precision (AP) values (%) of these two test sets with a slash ('/') separation.

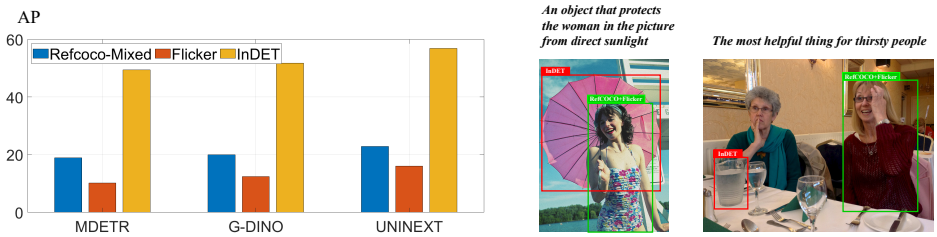
Method	Backbone	AP	AP by Group					
			G1	G2	G3	G4	G5	G6
MDETR	ResNet101	34.86 / 31.21	47.44 / 46.61	46.79 / 42.55	34.14 / 28.13	23.22 / 16.86	25.91 / 23.52	28.17 / 23.66
G-DINO	SwinB	35.96 / 30.43	47.10 / 45.91	47.17 / 42.56	35.29 / 27.28	26.84 / 18.46	27.95 / 23.74	27.61 / 23.57
UNINEXT	ResNet50	43.37 / 37.61	54.49 / 53.09	54.52 / 49.91	44.49 / 35.59	37.17 / 28.30	31.41 / 28.28	32.01 / 27.52
DROD (Ours)	ResNet50	62.24 / 53.78	67.14 / 65.08	67.34 / 61.56	60.89 / 48.82	55.10 / 41.50	70.15 / 64.64	74.22 / 67.11
DROD (Ours)	ViT-H	66.90 / 57.32	72.53 / 69.79	72.47 / 65.44	66.42 / 52.50	59.86 / 46.01	73.34 / 67.82	75.92 / 68.73

In our InDET test set, we compare our DROD model to other methods under the evaluation metric of object bbx average precision with a threshold of 0.5. On the other hand, we investigate whether these methods have truly comprehended the meaning of instruction, or they perform ROD only based on the key words (e.g., noun) without comprehending the whole expression. So we shuffle the InDET test set by randomly ordering the words in each instruction. We produce results of existing VG methods on our InDET test set without assuming object numbers in advance. For one method, if its performance drops more on the shuffled data, this method is shown to better comprehend the meaning of instruction.

Table 2 shows the evaluation results. Overall, UNINEXT achieves a higher AP than MDETR (i.e., 43.37 v.s. 34.86) in our InDET test set, while decreasing more than MDETR (i.e., 37.61 v.s. 31.21) in shuffled data. This indicates that UNINEXT is more effective than MDETR for ROD and better comprehends instruction meaning. Meanwhile, UNINEXT achieves a higher AP value than Grounding-DINO. In comparison, our DROD largely surpasses UNINEXT (62.24 v.s. 43.37) on the overall AP comparison, and using a VIT encoder further increases our performance. This indicates that our DROD is more effective to comprehend generalized instructions for ROD. Meanwhile, we observe that our performance drop is larger than UNINEXT (8.46 v.s. 5.76), which shows that our

Table 3: Evaluation results on the RefCOCO/g/+ datasets. We follow evaluation protocols to report AP values (%) of comparing methods. We use the notations "CC", "VG", "OI", "O365", "RIGame", for COCO, Visual Genome, OpenImage, Objects365, ReferItGame, respectively.

Method	Backbone	Data	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val-u	test-u
RefTR	ResNet101	VG	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
SeqTR	DarkNet53	VG,RIGame,Flickr,RefC	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37
MDETR	ResNet101	GoldG,CC,RefC	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
G-DINO	SwinB	O365,CC,RefC,GoldG,etc	83.95	87.79	79.16	72.91	80.91	62.96	76.98	76.76
UNINEXT	ResNet50	O365,CC,RefC	87.64	90.35	83.49	78.14	83.22	68.71	80.96	81.86
DROD (Ours)	ResNet50	O365,CC,InDET	88.92	90.86	85.57	78.27	83.39	71.04	83.01	82.91



(a) Numerical Results (b) Visual Comparisons

Figure 7: Our InDET dataset improves logic reasoning of ROD models. In (a), existing models trained with our InDET dataset show superior results compared to other datasets. In (b), we show visual comparisons by using the same DROD model but with different training datasets.

model better comprehends different expressions. Specifically for the results in each group, we notice that our performance drop is little in G1, and becomes larger from G2 to G4. This is because more and more words are introduced from G1 to G4 for object description. A random order gradually affects our model comprehension. For G5 and G6, we note that our method largely outperform other methods. The multi-object instructions incorporated in the dataset improves our performance.

Besides evaluating our InDET test set, we compare our DROD model with existing VG methods (Zhu et al., 2022; Li & Sigal, 2021) on the standard VG benchmarks RefCOCO/g/+. Table 3 shows the evaluation results. Overall, our DROD model achieves favorable performance on these datasets. This is because our DROD model utilizes InDET dataset where diversified instructions improve model generalizations. By using a conventional ROD model, we improve the VG performance from the data diversity perspective.

In addition to the overall precision comparisons, we evaluate how our dataset improves logic reasoning and instruction comprehension of existing models. Specifically, we select 2k test samples from our InDET test dataset where logic reasoning on instructions is required for object detection. For each model (i.e., MDETR, G-DINO, or UNINEXT), we train it by using different datasets (i.e., RefCOCO, Flicker, or InDET) and show the performance comparison on our 2k test samples. Fig. 7(a) shows the evaluation results where each model trained with our InDET dataset outperforms the same model trained with other datasets. In Fig. 7(b), we show visual comparisons by using our DROD model but with different training sets. It shows that using original datasets, the model tends to ground keywords rather than preform multi-modal reasoning based on instructions. In comparison, by training with our InDET, the model well interprets instruction meaning and conduct logic reasoning across languages and visual images.

7 Concluding Remarks

We aim to push ROD into practical usage from a data-centric perspective. On one hand, we notice that current REC expressions are insufficient to encompass user detection intentions. On the other hand, foundation models have shown promising generalizations to simulate manual understanding and description abilities. To this end, we develop InstructDET that leverages foundation models to produce human-like expressions in REC, which tends to incorporate common user intentions into ROD training. As a result, our DROD model achieves favorable performance compared to existing VG methods. In the future, we can combine our method with open-set object detectors to fully explore in-the-wild images (e.g., Internet images) for comprehensive user expression generation. We expect our DROD model to generalize as much as existing foundation models, and thus take a huge step towards completely solving ROD task.

ACKNOWLEDGEMENT

This paper is supported by the National Natural Science Foundation of China under Grants (62073245, 62173248,62233013). Shanghai Science and Technology Innovation Action Plan (22511104900). Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities.

REFERENCES

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B. Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. In *International Conference on Learning Representations*, 2021.

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

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning. *ArXiv preprint arXiv:2301.00234*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Special Interest Group on Knowledge Discovery and Data Mining*, 1996.

Shangqian Gao, Burak Uzkent, Yilin Shen, Heng Huang, and Hongxia Jin. Learning to jointly share and prune weights for grounding based vision and language models. In *International Conference on Learning Representations*, 2023.

Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo. Revitalizing cnn attentions via transformers in self-supervised visual representation learning. In *Advances in Neural Information Processing Systems*, 2021.

Chongjian Ge, Jiangliu Wang, Zhan Tong, Shoufa Chen, Yibing Song, and Ping Luo. Soft neighbors are positive supporters in contrastive visual representation learning. In *International Conference on Learning Representations*, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Drew Hudson and Christopher Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Gaoussou Youssouf Kebe, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Donald Engel, Edward Raff, Francis Ferraro, and Cynthia Matuszek. A spoken language dataset of descriptions for speech-based grounded language learning. In *Advances on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *ArXiv preprint arXiv:2304.02643*, 2023.

Noriyuki Kojima, Hadar Averbuch-Elor, and Yoav Artzi. A joint study of phrase grounding and task performance in vision and language models. *ArXiv preprint arXiv:2309.02691*, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *ArXiv preprint arXiv:2305.03726*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint arXiv:2301.12597*, 2023b.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In *Advances in Neural Information Processing Systems*, 2021.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv preprint arXiv:2304.08485*, 2023a.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ArXiv preprint arXiv:2303.05499*, 2023c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

OpenAI. Gpt-4 technical report. 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Bryan Plummer, Nikoli Dryden, Julius Frost, Torsten Hoefler, and Kate Saenko. Neural parameter allocation search. In *International Conference on Learning Representations*, 2022.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE/CVF International Conference on Computer Vision*, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *ArXiv preprint arXiv:2304.06712*, 2023.

Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv preprint arXiv:2302.13971*, 2023.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, 2022.

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv preprint arXiv:2303.04671*, 2023.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *ArXiv preprint arXiv:2306.04356*, 2023.

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

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *ArXiv preprint arXiv:2205.10625*, 2022.

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

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv preprint arXiv:2010.04159*, 2020.