Zero-shot Referring Expression Comprehension via Structural Similarity Between Images and Captions

Zeyu Han[1], Fangrui Zhu[1], Qianru Lao[2], Huaizu Jiang[1]
[1]Northeastern University [2]Harvard University
{han.zeyu,zhu.fang,h.jiang}@northeastern.edu, estherbear17@gmail.com

Abstract

Zero-shot referring expression comprehension *aims at localizing bounding boxes in an image corresponding to provided textual prompts, which requires: (i) a fine-grained disentanglement of complex visual scene and textual context, and (ii) a capacity to understand relationships among disentangled entities. Unfortunately, existing large vision-language alignment (VLA) models,* e.g.*, CLIP, struggle with both aspects so cannot be directly used for this task. To mitigate this gap, we leverage large foundation models to disentangle both images and texts into triplets in the format of* (`subject`, `predicate`, `object`)*. After that, grounding is accomplished by calculating the structural similarity matrix between visual and textual triplets with a VLA model, and subsequently propagate it to an instance-level similarity matrix. Furthermore, to equip VLA models with the ability of relationship understanding, we design a triplet-matching objective to fine-tune the VLA models on a collection of curated dataset containing abundant entity relationships. Experiments demonstrate that our visual grounding performance increase of up to 19.5% over the SOTA zero-shot model on RefCOCO/+/g. On the more challenging Who's Waldo dataset, our zero-shot approach achieves comparable accuracy to the fully supervised model. Code is available at https://github.com/Show-han/Zeroshot_REC.*

1. Introduction

Visual grounding is a fundamental task across computer vision and natural language processing, where the goal is to find the correspondences between image content and textual descriptions. It has broad applications in image captioning [18, 53], visual question answering [43, 66], vision-language navigation [11], etc. Collecting detailed grounding annotations to train specialist models, however, is cumbersome. Therefore, **zero-shot visual grounding** [30, 38, 54] is an attractive alternative.

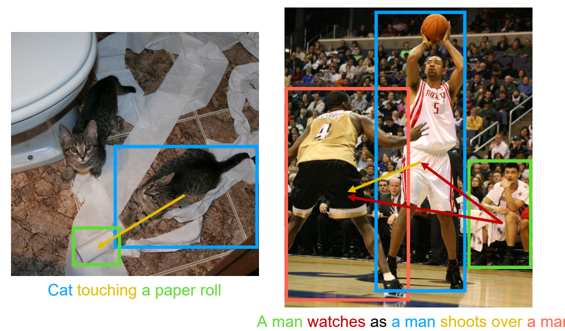

Figure 1. **Illustration of how we disambiguate visual entities based on their interactions with other entities.** The same entity or relationships in the image and caption are in the same color.

As a visual grounding task, the essence of referring expression comprehension (REC) is the alignment of text queries with corresponding image regions. To achieve this goal, it is critical for a grounding model to understand the relationships of entities[28], both within and cross different modalities (visual *vs.* textual), when identifying referred entities within an image. As shown in the upper part of Fig. 1, the `touching` relationship is the key to resolve the ambiguity of identifying the correct `cat`. Similarly, the lower part of Fig. 1 illustrates a more complex situation where multiple entities are engaged in various interactions. Both scenarios highlight the importance of *relationship understanding within both the image and caption*, where entities are not merely isolated elements but interact dynamically with others in the scene. In a zero-shot learning context, the task of understanding these relationships can be more challenging, as the model lacks exposure to specific training instances that could aid in interpretation.

Recent advances in zero-shot REC [30, 54, 65] have been largely driven by the integration of large-scale vision-language aligned (VLA) models such as CLIP [47] and FLAVA [52], which serve as bridges connecting text and image domain. These approaches, however, fall short in relationship understanding. On the one hand, in the textual domain, existing approaches adopt hand-crafted language parsers [54, 65] to decompose the input caption into a set of phrases, which are fragile and do not generalize well to long, complex captions in real-world applications, as shown in the bottom of Fig. 1. On the other hand, the visual relationship understanding capability of VLA models is inherently not good enough. Recent studies have revealed that VLA models behave like “bags-of-words” [67], and demonstrated that they fail to perform beyond chance level at simple tasks requiring compositional understanding [7, 8, 12, 57, 67]. Some effort have been dedicated to mitigate this issue by generating hard negative prompts through word replacement [8, 20, 67], caption augmentation [7, 8], and feature augmentation [22]. These rule-based methods, however, are limited in producing diverse samples and have potential bias of design pattern, consequently restricting their generalization capabilities.

In this paper, we focus on the REC task by explicitly modeling the entity relations within both images and captions using the *structural similarity* between them to solve the zero-shot visual grounding problem. Specifically, we decompose the image and caption into two sets of triplets in the form of (`subject`, `predicate`, `object`), in which each triplet captures a pair of potential entities with their interrelation. By considering the similarity of `subject`, `object`, and `predicate` jointly, we can find better matchings of object proposals and their referrings. Compared with existing work [54], our approach is more principled and eliminates the ad-hoc post-processing spatial relation resolver. More importantly, to improve the relationship understanding in the caption, we resort to ChatGPT and leverage its powerful *in-context learning* capability [56, 59] for the triplet decomposition to find all possible relation triplets given a sentence. In contrast to the dependency parser in [54], our parsing works better when dealing with long captions and do not restrict to spatial relationships (*e.g.*, `to the left of`), which can fully capture the rich compositional semantics present in actions or interactions, such as `walking` or `talking to`.

To address the limitation of a VLA model's visual relationship understanding, we harness a curated collection of data sources rich in relational knowledge, which include human-object interaction datasets [3, 46] and image scene graph dataset [27]. Similar to our grounding pipeline, we isolate visual entities and construct triplets in both visual and textual sides and then implement a triplet-level contrastive learning objective to fine-tune the VLA model. Compared with the existing rule-based negative prompts construction, this design has two unique advantages. First, by decomposing a single image into multiple triplets, we can obtain more training instances and improve the diversity of the training data than simply using the entire image for fine-tuning. Furthermore, the isolation of entities removes the distraction of other content in the image, providing more useful supervision for the model fine-tuning. We fine-tune a VLA model in a parameter-efficient manner [15] using LoRA [19], improving its visual relationship understanding while preserving its powerful generic feature representations learned from large-scale data. Our resulting model is called **VR-VLA** (**V**isual **R**elationship VLA).

Experimental results show that on the standard RefCOCO/g/+ datasets [42, 63], we can surpass the SOTA zero-shot baseline [54] up to 19.5%, and an average of 9.7%. On the challenging Who's Waldo dataset [6], whose captions are much longer and depict much richer interactions of humans, our zero-shot method significantly outperforms [54], achieving comparable accuracy to supervised methods. We also conduct ablation studies validating the effectiveness of our model design.

To summarize, our main contributions are three-folded. (1) A novel zero-shot visual grounding model, where we harness the powerful capabilities of foundation models (*e.g.*, ChatGPT and CLIP) to explicitly model the structural similarity between entities. (2) A novel recipe to improve a VLA model's visual relationship understanding by efficiently incorporating the supervision from a collection of curated data sources. (3) We report SOTA zero-shot visual grounding results on the REC datasets and also show promising results on the Who's Waldo dataset, where our zero-shot approach achieves comparable accuracy to the fully supervised method.

2. Related Work

Visual grounding. Based on the focus of the grounding task, visual grounding diverges into two main categories. The first emphasizes the noun properties of the query text. Precise understanding of the noun's meaning enables locating the corresponding grounding box in the image. Representative datasets contain MS-COCO [37], Object365 [49] (fixed-category), Flickr30K Entities [45] (open-vocabulary), etc. The second category emphasizes comprehending the interrelations among entities to localize the correct visual entity (potentially from among many similar ones) corresponding to the query text. Representative datasets contain RefCOCO/+/g [42, 63], Who's Waldo [6], etc. Our research of interest belongs to the second category, where relations are strong grounding clues.

Based on the use of grounding data, visual grounding can be classified into supervised and zero-shot categories. The majority of research [23, 35, 73] focuses on supervised visual grounding, where models are specifically designed and trained with grounding data for this purpose. Conversely, zero-shot visual grounding methods [30, 38, 54] adapt pre-existing vision-language models for grounding tasks. Our work is situated within this zero-shot visual grounding paradigm.

Visual relationship understanding. Images and texts are constructed using fundamental elements — objects in images and noun spans in texts — along with their interactions and relationships. For a model to understand relationship, it must not only detect individual entities but also establish relational links between them. Research communities such as human-object interaction [3, 10, 24, 25, 34, 40, 46, 69–71, 75] and visual scene graph [27, 33, 55, 60, 61, 68, 72, 76] emphasize the relational aspect for visual tasks.

Visual relationship understanding is also highly relevant to the *compositional reasoning* [8, 12, 17, 20, 22, 67] ability for VLA models. Although we anticipate that VLA models, trained via contrastive learning on extensive image-text pairs, would inherently develop a capacity for compositional reasoning, the reality is somewhat different. Most SOTA VLA models behave like "bags-of-words" [67]. They are capable of matching textual entities with corresponding visual elements, but falling short in interpreting their relationships or attributes. To address this, many studies have implemented strategies such as introducing negative text prompts [7, 8, 67] in training batches — including noun substitutions, attribute and verb modifications, caption augmentation — or integrating scene graphs [17] during training or inference. In our work, we show that fine-tune the VLA models on the visual relationship datasets can alleviate this problem. Furthermore, the explicitly defined structural representation also help strengthen the compositional reasoning for visual grounding tasks.

Language parsing. In the language side, structure prediction [1, 5, 44, 50] is well studied and aims for solving several problems including entity recognition [9, 32, 51], relation classification [14, 74], semantic role labeling [2, 16], event extraction [21, 31], coreference resolution [29, 58], etc. The acquired structural representation can be illustrated as either a language parsing tree [54] or a set of labels that indicate the respective roles of each word [50].

3. Proposed Approach

Our proposed grounding pipeline contains two stages. First, we decouple image and text entities and construct triplets in the format of $(\texttt{subject}, \texttt{predicate}, \texttt{object})$. Second, we calculate triplet-level similarity matrix and propagate it to the instance-level and then obtain the bounding box with the highest similarity score. The primary focus in our matching pipeline is to accurately model the relationship between entities, which is achieved by the the triplet-level structural similarity, as shown in Fig. 2. We also provide a novel recipe to equip the VLA models with better compositional understanding ability.

3.1. Constructing Triplets

Given a caption, denoted by $\mathcal{C}$, and its associated image, denoted by $\mathcal{I}$, we postulate that both $\mathcal{C}$ and $\mathcal{I}$ comprise sets

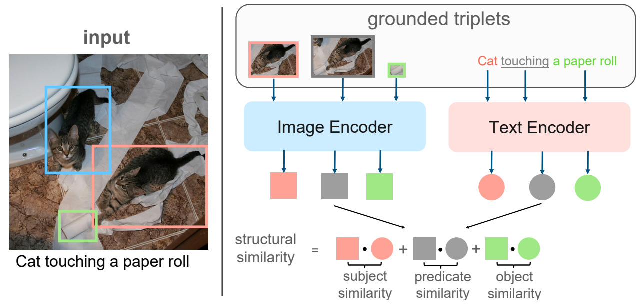

Figure 2. **Illustration of the triplet-level structural similarity.** Visual and textual triplets are encoded by image encoder and text encoder, respectively. Then the structural similarity is calculated as the sum of cosine similarities between `subject`, `predicate`, and `object`.

of entities[1], denoted by $\mathcal{E}_T = \{e_i^T\}_{i=1}^M$ for the text and $\mathcal{E}_I = \{e_i^I\}_{i=1}^N$ for the image, where M and N represent the total number of entities in $\mathcal{C}$ and $\mathcal{I}$, respectively. Interrelation of an entity pair is represented by $r^T(\cdot)$ for the text and $r^I(\cdot)$ for the image. In this stage, our objective is to construct entity-relation triplets for both modalities.

For text, we denote the triplets as:

$$\mathcal{T}_T = \{t_{ij}^T = (e_i^T, r^T(e_i^T, e_j^T), e_j^T) \,|\, 1 \leq k, l \leq M\}. \quad (1)$$

For image, we denote the triplets as:

$$\mathcal{T}_I = \{t_{kl}^I = (e_k^I, r^I(e_k^I, e_l^I), e_l^I) \,|\, 1 \leq i, j \leq N\}. \quad (2)$$

The cardinality of the above two sets are defined as M' for text and N' for image, respectively.

Textual triplets construction. Large language models have exhibited a powerful capacity for a range of downstream tasks. Here, we leverage its powerful in-context learning capability to parse a caption $\mathcal{C}$ into triplets $\mathcal{T}_T$. Specifically, We design a prompt to instruct the ChatGPT to parse the caption text $\mathcal{C}$. Fig. 3 provides an overview on how we design prompt for RefCOCO/+/g dataset, and further details are elaborated as follows. Note that the prompts can vary depending on datasets to accommodate different distributions of the data.

As shown in Fig. 3, the prompt can be divided into four parts: *(i) general instruction*; *(ii) supporting details*; *(iii) in-context learning examples*, and *(iv) task instruction*, followed by *LLM completion*, which yields the output of the LLM in the specified format and content. In part (i), we define a clear and general instruction for specific task. Then, we elaborate supporting details in part (ii), including the expected output format, essential elements, and preferences for what should or should not be included, etc. In part (iii), we curate several in-context learning examples to guide the

[1] We use "entities" to denote objects to differentiate from `object` in a triplet.

General Instruction	`Given a sentence, first determine the main entity (with its attributes) that the sentence is describing. Then analyze and extract all the spacial relation/action between the determined main entity with other entities.`
Supporting Details	`Return in the JSON format: {"entity": "xxx", "relations": [["xxx", "yy", "zz"], ["aa", "bb", "xxx"], ...]}. All the returned relations must be a triplet containing exactly three elements. The first… The second… The third… You can make some guess if… Your response must accurately follow the above instruction in both content and format, akin to the examples provided below, without any extra explanatory text.`
In-context Learning Examples	`##Examples##` *INPUT: a woman wearing blue jeans sitting on a chair with a baby sitting in her lap* *OUTPUT: {"entity": "a woman", "relations": [["a woman", "wearing", "blue jeans"], ["a woman", "sitting on", "a chair"], ["a baby", "sitting in", "a woman"]]}* *INPUT: ......* *OUTPUT: ......*
Task Instruction	`##Your Task##` *INPUT:* [user input, e.g., one small girl in white t-shirt is touching the elephant] *OUTPUT:*
LLM completion	*{"entity": "one small girl", "relations": [["one small girl", "wearing", "white t-shirt"], ["one small girl", "is touching", "the elephant"]]}*

Figure 3. **Illustration of leveraging ChatGPT's powerful in-context learning capability to parse a caption into triplets.**

LLM, which is immediately followed by part (iv), where we append the input caption $\mathcal{T}$. Finally, we feed the above input into LLM, and then decoupled textual triplets will be generated through the LLM completion. We also do a simple format check after the completion.

Visual triplets construction. In an image, entities are represented by bounding boxes, each enclosing an individual object. These boxes might be predefined by the dataset or extracted using a pre-trained object detector. Without prior knowledge about how these entities are related, we assume potential interactions can happen between every pair of entities. Therefore, we generate visual pairs using a Cartesian product, which includes all possible combinations of entities. A notable case is when a pair consists of the same entity (box) twice. This represents a self-relation, suggesting the entity's own attributes, such as color (*e.g.*, `red`) or self-actions (*e.g.*, `walking`). Then we use the union area of two entity boxes to represent $r^I(e_i^I, e_j^I)$, the interrelation between entities.

Here, we derive triplets set $\mathcal{T}_T$ and $\mathcal{T}_I$ from caption and image, respectively. Before moving on, we filter out redundant triplets $t_{ij}^I \in \mathcal{T}_I$ based on heuristic rules similar to ReCLIP [54]. Specifically, given a textual triplet t_{ij}^T, where its `predicate` contains keywords that reflect some spacial relationships, *e.g.*, `to the left of`. In such a case, we filter out visual box pairs where the central point of the former box (*i.e.*, `subject`) is in the right of the latter one (*i.e.*, `object`). This approach is much simpler than building complicated spatial semantic trees like in ReCLIP, yet it effectively adds spatial context and improves performance.

3.2. Grounding Based on the Structural Similarity

With the triplets derived from the image and caption, we can use them to solve the visual grounding problem, which allows us to leverage the structural similarity between both modalities to more accurately link the textual descriptions of their corresponding image regions.

We consider two grounding directions: `text` $\rightarrow$ `image` and `image` $\rightarrow$ `text`, each serving unique task requirements. The `text` $\rightarrow$ `image` grounding is applied when identifying specific image regions based on a textual description (*e.g.*, RefCOCO/+/g [42, 63] dataset). Conversely, `image` $\rightarrow$ `text` grounding involves locating relevant textual descriptions for a given image region (*e.g.*, Who's Waldo [6] dataset). Given their symmetrical nature, this section will primarily focus on the `text` $\rightarrow$ `image` grounding scenario.

Triplet-level grounding. Given a text triplet $t_{ij}^T = (e_i^T, r^T(e_i^T, e_j^T), e_j^T)$, we separately feed e_i^T, $r^T(e_i^T, e_j^T)$, and e_j^T into a VLA text encoder to obtain three text embeddings, denoted as $(\mathbf{t}_i, \mathbf{t}_{i,j}, \mathbf{t}_j)$. Similarly, for a image triplet $t_{kl}^I = (e_k^I, r^I(e_k^I, e_l^I), e_l^I)$, we derive three image embeddings $(\mathbf{v}_k, \mathbf{v}_{k,l}, \mathbf{v}_l)$ using the image encoder. The similarity between these two triplets is then given by:

$$\mathbf{S}(t_{ij}^T, t_{kl}^I) = \cos(\mathbf{t}_i, \mathbf{v}_k) + \cos(\mathbf{t}_{i,j}, \mathbf{v}_{k,l}) + \cos(\mathbf{t}_j, \mathbf{v}_l), \quad (3)$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity function. $\mathbf{S} \in \mathbb{R}^{M' \times N'}$ is the similarity matrix between all text triplets with all image triplets. We subsequently get a binary indicator matrix $\mathbf{B} \in \{0, 1\}^{M' \times N'}$ by:

$$\mathbf{B}(t_{ij}^T, t_{kl}^I) = \begin{cases} 1 & \text{if } k, l = \arg\max_{m,n}(\mathbf{S}(t_{ij}^T, t_{mn}^I)). \\ 0 & \text{otherwise.} \end{cases}$$

Here, for each text triplet t_{ij}^T, the binary indicator matrix $\mathbf{B}$ assigns the value of 1 to the most similar image triplet t_{kl}^I and 0 to all others.

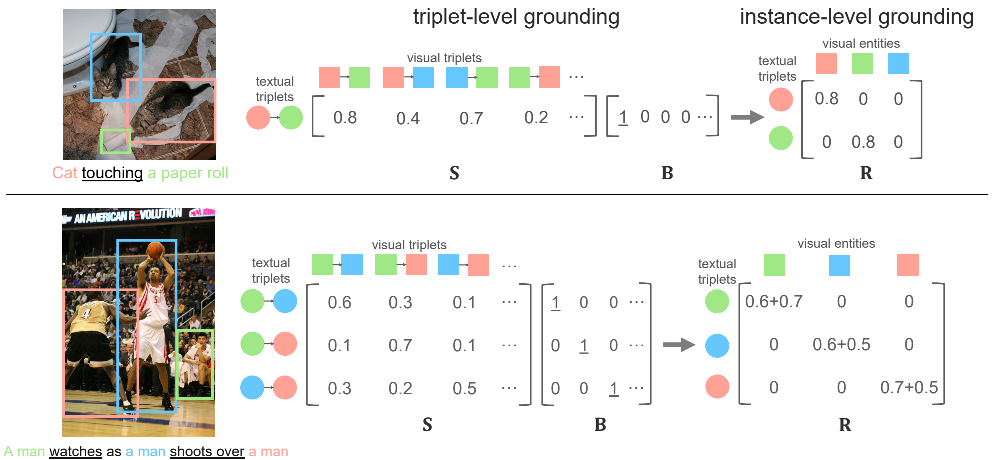

Figure 4. **Illustration of propagating the similarity scores from grounded triplets to the instance level.** Via the aggregation of the similarity scores from multiple grounded triplets, it helps find the instance-level correspondences more accurately. For instance, in the lower part, the referring expression a man and the blue bounding box appear in two different triplets, acting as the `subject` and `object`, respectively. Such structural similarity provide more useful cues to improve the instance-level grounding. (Best viewed in color.)

Instance-level grounding. Another substantial challenge in the visual grounding problem is that both the `subject` and `object` in a triplet may have multiple interactions with other entities. To this end, we design a novel method to propagate the triplet-level grounding results to the instances.

Specifically, based on the triplet-level grounding results, we can compute the instance-level structure-aware similarity matrix $\mathbf{R}$ as follows:

$$\mathbf{R}(e_i^T, e_k^I) = \sum_{j,l} \mathbf{B}(t_{ij}^T, t_{kl}^I)\mathbf{S}(t_{ij}^T, t_{kl}^I) + \sum_{j,l} \mathbf{B}(t_{ji}^T, t_{lk}^I)\mathbf{S}(t_{ji}^T, t_{lk}^I). \tag{4}$$

The two terms in Eq. 4 consider both the cases where e_i^T and e_k^I appear as the `subject` and `object` in different triplets, as shown in lower part of Fig. 4. By aggregating the similarity scores from multiple grounded triplets, it helps find the instance-level correspondences more accurately.

Finally, for each text entity, we compute the most relevant image entity as follows:

$$\hat{e}_k^I = \mathrm{argmax}_m \mathbf{R}(e_i^T, e_m^I), \tag{5}$$

where $\hat{e}_i^I$ denotes the corresponding visual entity for e_i^T. Notably, our approach can easily extend to the one-to-many grounding scenario if we implement a threshold-based selection in place of the `argmax` function in Eq. 4 and Eq. 5. We limit our discussion to one-to-one grounding for clearer understanding.

3.3. Enhanced Relational Understanding

In Equation 3, the interaction between two entities is represented by the term $\cos(\mathbf{t}_{i,j}, \mathbf{v}_{k,l})$. This term is crucial as it attempts to quantify the relationship between entities through cosine similarity, under the assumption that VLA models can adequately grasp these relationships. Nevertheless, as indicated by other studies [54, 67], this assumption often falls short in practice. To address this issue, we fine-tuned VLA models using a combination of datasets rich in relational knowledge. These datasets include HICO-det [3], SWiG [46], and Visual Genome (VG) [27]. Notably, in the case of the VG dataset, we excluded all the images from COCO to maintain the integrity of the zero-shot protocol, aligning with our experiments based on RefCOCO/+/g.

The datasets mentioned provide annotation bounding boxes for objects together with their textual descriptions and relationships with other objects. So we can easily follow what we have done in the triplet-level grounding stage to create visual-textual triplets, and then utilize a contrastive learning loss on these triplets. To clarify, we use the same notation in Eq. 3 to calculate the similarity between two triplets. Assume t_{ij}^T and t_{kl}^I are two corresponding triplets, we define the contrastive loss as follows:

$$\mathcal{L} = \sum_{(t_{ij}^T, t_{kl}^I)} \left[\log\left(\frac{\mathbf{S}(t_{ij}^T, t_{kl}^I)}{\sum_{m,n} \mathbf{S}(t_{ij}^T, t_{mn}^I)} \right) + \log\left(\frac{\mathbf{S}(t_{ij}^T, t_{kl}^I)}{\sum_{m,n} \mathbf{S}(t_{mn}^T, t_{kl}^I)} \right) \right]. \tag{6}$$

Through this refined approach, we aim to enhance the VLA models' ability to understand and accurately score the relationship between entities, thereby enhancing the zero-shot grounding capability.

Compared with the existing rule-based negative prompts construction [7, 8, 67], this design has two unique advan-

Model	RefCOCOg Val	RefCOCOg Test	RefCOCO+ Val	RefCOCO+ TestA	RefCOCO+ TestB	RefCOCO Val	RefCOCO TestA	RefCOCO TestB
Random	18.12	19.10	16.29	13.57	19.60	15.73	13.51	19.20
Supervised SOTA [35]	88.73	89.37	85.24	89.63	79.79	92.64	94.33	91.46
CPT-Blk w/ VinVL [62]	32.10	32.30	25.40	25.00	27.00	26.90	27.50	27.40
CPT-Seg w/ VinVL [62]	36.70	36.50	31.90	35.20	28.80	32.20	36.10	30.30
CLIP (ViT-B/32)								
CPT-adapted [54]	21.77	22.78	23.46	21.73	26.32	23.79	22.87	26.03
GradCAM [48]	49.51	48.53	44.64	50.73	39.01	42.29	49.04	36.68
ReCLIP [54]	56.96	56.15	45.34	48.45	42.71	45.77	46.99	45.24
Ours	57.60	56.64	45.64	47.59	42.79	48.24	48.40	49.15
Ours+VR-CLIP	59.87	59.90	**55.52**	**62.56**	45.69	**60.62**	**66.52**	**54.86**
FLAVA								
Ours	<u>60.95</u>	<u>59.99</u>	48.89	50.02	<u>46.86</u>	49.37	47.76	51.68
Ours+VR-FLAVA	**61.25**	**60.86**	<u>50.79</u>	<u>53.35</u>	**47.62**	<u>52.46</u>	<u>52.66</u>	<u>52.92</u>

Table 1. **Accuracy on the RefCOCOg, RefCOCO+ and RefCOCO datasets.** `Ours` represents leveraging our triplet-to-instance pipeline for grounding. `Ours+VR-CLIP/VR-FLAVA` further replaces the original VLA model with our relationship-enhanced model. Except for the supervised method, the best results are highlighted in **bold**, and second-best results are <u>underlined</u>.

tages. First, by decomposing a single image into multiple triplets, we can obtain more training instances and improve the diversity of the training data than simply using the entire image for fine-tuning. Furthermore, the isolation of entities removes the distraction of other content in the image, providing more useful supervision for the model fine-tuning.

4. Experiments

4.1. Setup

RefCOCO/RefCOCO+[42]/RefCOCOg[63] are collected from MS-COCO [36]. RefCOCO includes 19,994 images with 142,210 referring expressions. RefCOCO+ has 19,992 images and 141,564 expressions. RefCOCOg contains 26,771 images with 104,560 expressions. In RefCOCO and RefCOCO+, expressions are shorter, averaging 1.6 nouns and 3.6 words. In RefCOCOg, expressions are longer, averaging 2.8 nouns and 8.4 words.

Who's Waldo [6] introduces a person-centric visual grounding task, where all names in the captions are masked, forcing models to link boxes and the masked [NAME] tokens through attributes and interactions between visual entities. The captions are long and contain complex scene descriptions. We use its test split for evaluation, which contains 6741 images. Each caption contains about 30 words.

Evaluation metrics. On RefCOCO/+/g, we follow previous work [54] to use accuracy as the grounding results, *i.e.*, if the IoU (Intersection over Union) value between the predicted box and ground truth region is larger than 0.5, it is a correct prediction. On Who's Waldo, following previous work [6], given the grounding results of person in textual descriptions to bounding boxes in the image, we report the accuracy against ground-truth links on the test set.

4.2. Implementation Details

RefCOCO/+/g On RefCOCO/+/g datasets, we adopt the test-time augmentation strategy outlined in ReCLIP [54], where they use both cropping and blurring for isolated visual entities. When blurring the union region, we separately process each box in the box pair, isolating only the area where the box is located, rather than directly using the union area, which helps to minimize distraction from other visual objects. We use the whole caption instead of the decoupled main entities before feeding into the text encoder since it produces better results. All the heuristic rules mentioned in Sec. 3.1 are also adopted from ReCLIP.

VLA fine-tuning We fine-tune the CLIP model based on the code from [8], where LoRA rank $r = 4$, batch size is 1024, learning rate is $5e-6$, and epoch is 20. We use the huggingface PEFT [41] for fine-tuning FLAVA, where we set lora rank $r = 16$ for 10 epochs. Since the SWiG dataset [46] does not contains triplet annotations, we use ChatGPT to convert the existing annotation into triplets.

Box generator Following ReCLIP, we use bounding boxes generated from MAttNet [64] as the box proposals on RefCOCO/+/g. On the Who's Waldo dataset, we use the box proposals provided in the annotations.

4.3. Main Results

RefCOCO/+/g We benchmark our approach against various zero-shot visual grounding models, including Colorful Prompt Tuning (CPT) [62], GradCAM [48], and ReCLIP [54]. CPT-adapted is introduced and adapted by [54]. ReCLIP represents the latest SOTA in zero-shot REC methods.

As shown in Table 1, compared to other models utilizing the same CLIP architecture, our proposed method consis-

Method	Test Accuracy
Supervised	
Who's Waldo [6]	**63.5**
Pretrained on grounding data	
Gupta et al. [13] (COCO)	35.6
Gupta et al. [13] (Flickr30K)	38.2
SL-CCRF [39]	46.4
MAttNet [64]	24.1
UNITER [4]	34.2
CLIP	
ReCLIP [54]	29.4
Ours	60.8
Ours+VR-CLIP	<u>61.3</u>
FLAVA	
Ours	59.6
Ours+VR-FLAVA	59.8

Table 2. **Accuracy on the Who's Waldo dataset.** The best results are highlighted in **bold**, and second-best results are <u>underlined</u>.

tently outperforms all competitors across all splits. Specifically, our model exhibits a performance improvement of up to 19.53% over ReCLIP, with an average enhancement of 9.74%. Remarkably, even without fine-tuning the backbone CLIP model, our method can surpass the ReCLIP by up to 3.91%, with an average of 1.05%, showing that the structural similarity based on ChatGPT's parsing also contributes to the relational understanding.

In addition, we also extend our methodology to another VLA model, FLAVA [52], to verify the generalizability of our approach. Not surprisingly, when integrated into our matching pipeline, FLAVA demonstrates superior performance compared to the CLIP model. This can be attributed to FLAVA's inherently more robust architecture. After fine-tune the FLAVA, the resulting VR-FLAVA consistently improve the performance across all dataset splits, reinforcing the effectiveness of our method in enhancing the relationship understanding of various VLA models.

Who's Waldo We compare our approach with the models trained on grounding dataset, which include Gupta et al. [13], SL-CCRF [39], MAttNet [64] and UNITER [4]. The *Who's Waldo* method serves as a supervised baseline, as reported in its original paper [6]. Additionally, we adapte ReCLIP [54] for our dataset, utilizing their language parser to identify potential referring expressions, followed by grounding using their original approach[2].

As shown in Table 2, our approach outperforms all models trained on the grounding dataset, with a notable margin. Among these, SL-CCRF was the closest competitor, yet it falls behind our method by 14.9%. When compared to the supervised *Who's Waldo* method, our zero-shot setting only shows a 2.2% reduction in performance. This highlights

[2]The experiment is conducted using the authors' released code.

Triplet	VR-CLIP	Val	Test
✗	✗	55.35	54.33
✗	✓	56.90	56.81
✓	✗	57.60	56.64
✓	✓	**59.87**	**59.90**

Table 3. **Effectiveness of each component of our grounding pipeline on RefCOCOg.** Triplet means whether we use the triplet-to-instance grounding (instead of scoring-and-ranking). VR-CLIP represents whether we use fine-tuned VR-CLIP instead of the original CLIP model.

our method's effectiveness in visual grounding, especially in processing long and complex captions. Notably, performance of ReCLIP is no better than random choice. This is because their language parser fails to handle the real-world complex captions as in Who's Waldo. It validates our design choice of using the in-context learning capability of LLMs for better generalization ability.

We show visualization results in Fig. 5. The left two columns display results from RefCOCOg, while the rightmost column shows results from Who's Waldo. ReCLIP failed in all examples listed. Our approach, which explicitly models relationships (indicated by arrows in the images), provides more helpful information for grounding. Additionally, it is observed that ChatGPT consistently excels in parsing complex captions, such as those in Who's Waldo.

4.4. Ablation Studies

In this section, we conduct ablation studies on RefCOCOg. This dataset is particularly suitable for our evaluation because it has longer captions and rich entity interactions, making it an ideal testbed for assessing each component.

Effectiveness of components in the grounding pipeline. We explore two key variations: *Triplet* and *VR-CLIP*. The *Triplet* variation examines the impact of utilizing triplet-to-instance matching as opposed to a basic scoring-and-ranking approach, i.e., scoring each isolated boxes using CLIP, than select one with the highest similarity. The *VR-CLIP* variant assesses the performance differences between the fine-tuned VR-CLIP and the original CLIP model.

As shown in Table 3, substituting the CLIP model with the VR-CLIP yields superior results because of the enhanced relational capability. Note that although we do not explicitly use the structural similarity in this context, we are still using the whole caption as the referring expressions, which inherently convey the relationship information. Further analysis reveals that if we replace the scoring-and-ranking with our proposed triplet-to-instance matching pipeline, we can get better results through the relationship modeling. By combining both, we can achieve best results.

Effectiveness of triplet components. We separately remove the `subject`, `object` and `predicate` terms in

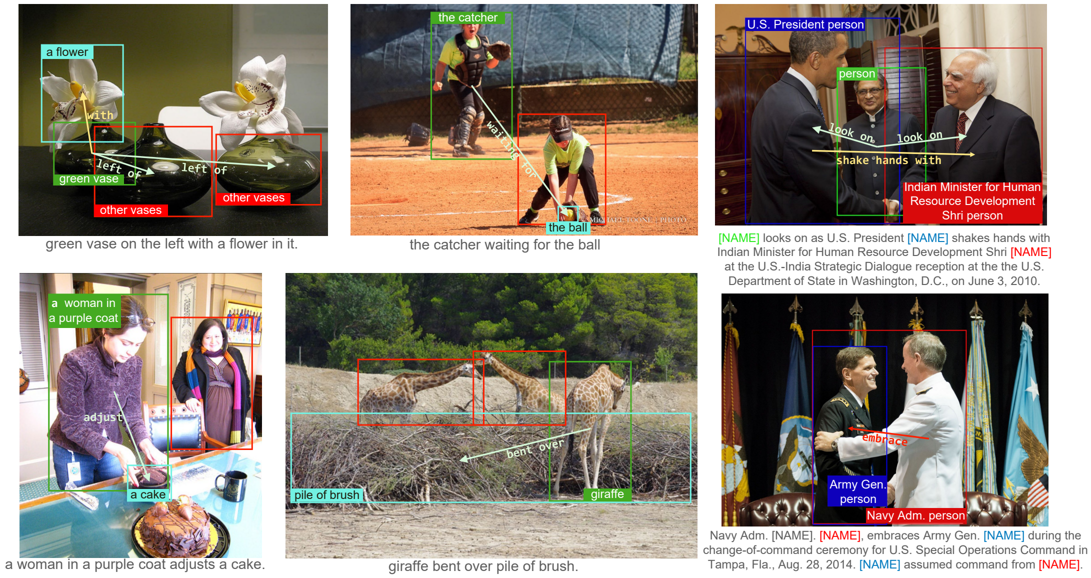

Figure 5. **Zero-shot visual grounding results.** Left two columns are results from RefCOCO, where our predictions are in green box, distraction objects are in red box. The rightmost column shows results from Who's Waldo, where predicted annotation links are in the same color. Arrows represent relationships between visual objects, and the text on the images are the parsed triplets.

	Val	**Test**
full	**59.87**	**59.90**
w/o `subject`	48.35	47.83
w/o `object`	56.92	57.05
w/o `predicate`	56.90	56.81

Table 4. **Effectiveness of each triplet component on RefCOCOg.** Each removal is done by set the corresponding term in Eq. 3 to 0. For w/o `predicate`, we also turn off the box pair filter to remove any spatial information provided in the caption.

Eq. 3 to explore their effectiveness in grounding performance. As shown in Table 4, the absence of `subject` or `object` leads to the loss of most noun information and its attributes, resulting in a clear accuracy drop. This impact is particularly significant for `subject` removal, since most main entities in RefCOCOg are represented as `subject`. Without `predicate`, meaning no interrelation between entities is considered, the accuracy degrades by about 3 points. It emphasizes the importance of our structural similarity in modeling the entity relationships.

Effectiveness of triplet- to instance-level grounding . We separately remove the first and second terms in Eq. 4 to validate our design of triplet- to instance-level grounding. The results are reported in Table 5, which shows that utilizing both pieces of information yields the best performance.

4.5. Limitations

While our method enhances the visual relationship understanding of VLA models, it sacrifices the model's zero-shot capabilities as generalist models and downgrades them to specialist ones. For example, after fine-tuning, CLIP's zero-shot image classification accuracy decreases from 0.63 to 0.50 and the R@5 for image retrieval on COCO decreases from 0.57 to 0.54. A promising future direction is to apply large-scale unlabeled image-caption pairs during VLA fine-tuning, and probably we can generate region-text triplets using ChatGPT and SAM (Segment Anything Model) [26] as pseudo ground-truths.

	Val	**Test**
full	**59.87**	**59.90**
w/o 1st term	59.54	59.37
w/o 2nd term	59.11	59.26

Table 5. **Effectiveness of different terms in Eq. 4 for triplet- to instance-level grounding on RefCOCOg.**

5. Conclusion

In this paper, we proposed a novel zero-shot referring expression comprehension model by resorting to powerful capabilities of foundation models (*e.g.*, ChatGPT and CLIP) to explicitly model the structural similarity between entities and then find their correspondences by propagating the similarity from triplets to instances. We also introduced a novel recipe to improve a VLA model's visual relationship understanding by training from a collection of curated data sources. Experimental results on RefCOCO/+/g and Who's Waldo validate the effectiveness of our approach.

References

[1] David Belanger and Andrew McCallum. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992. PMLR, 2016. 3

[2] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pages 152–164, 2005. 3

[3] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015. 2, 3, 5

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 7

[5] Despina Christou and Grigorios Tsoumakas. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, 9:62574–62582, 2021. 3

[6] Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021. 2, 4, 6, 7

[7] Sivan Doveh, Assaf Arbelle, Sivan Harary, Amit Alfassy, Roei Herzig, Donghyun Kim, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *arXiv preprint arXiv:2305.19595*, 2023. 2, 3, 5

[8] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 2, 3, 5, 6

[9] Jenny Rose Finkel and Christopher D Manning. Nested named entity recognition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 141–150, 2009. 3

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 3

[11] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022. 1

[12] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 2, 3

[13] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 7

[14] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*, 2018. 3

[15] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. 2

[16] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, 2017. 3

[17] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023. 3

[18] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[20] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. Structure-clip: Enhance multi-modal language representations with structure knowledge. *arXiv preprint arXiv:2305.06152*, 2023. 2, 3

[21] Heng Ji and Ralph Grishman. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: Hlt*, pages 254–262, 2008. 3

[22] Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. Comclip: Training-free compositional image and text matching. *arXiv preprint arXiv:2211.13854*, 2022. 2, 3

[23] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[24] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 3

[25] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *CVPR*, 2023. 3

[26] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 8

[27] 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*, 123:32–73, 2017. 2, 3, 5

[28] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6867–6876, 2018. 1

[29] Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*, 2018. 3

[30] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022. 1, 2

[31] Qi Li, Heng Ji, and Liang Huang. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, 2013. 3

[32] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*, 2019. 3

[33] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017. 3

[34] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*, 2022. 3

[35] Fangjian Lin, Jianlong Yuan, Sitong Wu, Fan Wang, and Zhibin Wang. Uninext: Exploring a unified architecture for vision recognition. *arXiv preprint arXiv:2304.13700*, 2023. 2, 6

[36] 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 *ECCV*, 2014. 6

[37] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[38] Xuyang Liu, Siteng Huang, Yachen Kang, Honggang Chen, and Donglin Wang. Vgdiffzero: Text-to-image diffusion models can be zero-shot visual grounders. *arXiv preprint arXiv:2309.01141*, 2023. 1, 2

[39] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 7

[40] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. Fgahoi: Fine-grained anchors for human-object interaction detection. *arXiv preprint arXiv:2301.04019*, 2023. 3

[41] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022. 6

[42] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 4, 6

[43] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1

[44] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *ICLR*, 2021. 3

[45] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[46] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020. 2, 3, 5, 6

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[48] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[49] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[50] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019. 3

[51] Takashi Shibuya and Eduard Hovy. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620, 2020. 3

[52] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 15638–15650, 2022. 1, 7

[53] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022. 1

[54] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, 2022. 1, 2, 3, 4, 5, 6, 7

[55] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 3

[56] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. 2

[57] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 2

[58] Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, 2020. 3

[59] Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, and Dongkuan Xu. Gentopia: A collaborative platform for tool-augmented llms. *arXiv preprint arXiv:2308.04030*, 2023. 2

[60] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 3

[61] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 3

[62] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 6

[63] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 4, 6

[64] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 6, 7

[65] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19456–19465, 2023. 1, 2

[66] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 1

[67] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 5

[68] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 3

[69] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *NeurIPS*, 2021. 3

[70] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021.

[71] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022. 3

[72] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*, 2017. 3

[73] 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. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 2

[74] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*, 2017. 3

[75] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, 2022. 3

[76] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *ICCV*, 2021. 3