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

Supplementary Material

A. Implementation Details

Special Cases for Textual Triplets. Not all caption can be perfectly decoupled into textual triplets. Often, such a caption is just a single noun or lacks an explicit subject. For example, the caption `red apple` would be decoupled into (`"red apple"`, `""`, `""`), and `person walking` into (`"person"`, `"walking"`, `""`). In these instances, we fill the blank spaces (*i.e.*, missing items in the triplet) with the `subject` string, resulting in (`"red apple"`, `"red apple"`, `"red apple"`) and (`"person"`, `"walking"`, `"person"`), respectively, for the previous examples. This approach ensures our grounding pipeline can manage such simplified cases. For instance, (`"red apple"`, `"red apple"`, `"red apple"`) will be matched three times with the visual entity of a red apple, which means that it degenerates to the naive score-and-ranking strategy.

Additionally, before feeding the `predicate` into the text encoder, on the RefCOCO/+/g dataset, we form a complete sentence by concatenating the `subject`, `predicate` and `object`, *e.g.*, `"vase on top of table"` instead of `"on top of"`. On the Who's Waldo dataset, we add `a person` before and after the `predicate`, *e.g.*, `"a person looking at a person"` instead of `"looking at"`. This is because, in most cases, a single `predicate` like `"on top of"` is semantically meaningless. Instead, a complete phrase like `"vase on top of table"` offers more contextual information.

Dataset for VLA Fine-tuning. Our dataset for VLA fine-tuning is obtained from HICO-det [1], SWiG [4] and Visual Genome (without COCO images) [2]. Each datapoint in this dataset consists of multiple image-text triplet pairs with the same text triplet, with two examples illustrated in Fig. 1. Notably, "multiple" is because we group image triplets corresponding to the same textual triplet into a single datapoint. Consequently, a single datapoint may comprise several distinct image triplets paired with the same textual triplet. For training purposes, we randomly select one image triplet from each category per epoch. This strategy is adopted to avoid scenarios in a single batch where an image triplet is forced to simultaneously pull in and push away from the same textual triplet due to the contrastive learning objective.

Figure 1. **Two examples in the dataset for VLA fine-tuning.**

B. Additional Experiment Results

This section provides additional experiment results tested on RefCOCO/+/g. Table. 1 shows the full results with different box proposal variants, *i.e.*, using a bounding box size prior (filter our objects smaller than 5% of the image), and use the groundtruth bounding boxes as box proposals.

C. Additional Visualization Results

In this section, we provide additional visualization results for RefCOCO, RefCOCO+, RefCOCOg, and Who's Waldo, illustrated in Fig.2, Fig.3, Fig.4, and Fig.5, respectively. Fig. 2, 3, 4 highlight examples where ReCLIP failed but our grounding approach yielded correct results. The textual triplets parsed by ChatGPT are also displayed in the images. Next, we will discuss some selected examples to illustrate the advantages of our approach.

As shown in Fig. 4, on the RefCOCOg dataset, our approach is able to successfully differentiate multiple instances of the same object category by understanding their relationships with others. For instance, in the first example in Fig. 4, the blue suitcase can be accurately grounded among other ones. For the example of "a zebra that is standing", the curved arrow in image represents self-action,

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
Random (w/ groundtruth box proposal)	20.18	20.34	16.73	12.57	22.13	16.37	12.45	21.32
Supervised SOTA [3]	88.73	89.37	85.24	89.63	79.79	92.64	94.33	91.46
CPT-Blk w/ VinVL [7]	32.10	32.30	25.40	25.00	27.00	26.90	27.50	27.40
CPT-Seg w/ VinVL [7]	36.70	36.50	31.90	35.20	28.80	32.20	36.10	30.30
CLIP								
CPT-adapted [6]	21.77	22.78	23.46	21.73	26.32	23.79	22.87	26.03
GradCAM [5]	49.51	48.53	44.64	50.73	39.01	42.29	49.04	36.68
ReCLIP [6]	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**
CLIP (w/ box size prior)								
CPT-adapted [6]	28.98	30.14	26.64	25.13	27.27	26.08	25.38	28.03
GradCAM [5]	52.29	51.28	49.41	59.66	38.62	44.65	53.49	36.19
ReCLIP [6]	**60.85**	**61.05**	55.07	60.47	47.41	54.04	58.60	49.54
Ours	58.52	57.95	52.38	57.65	45.65	56.10	58.97	52.23
Ours+VR-CLIP	58.95	59.55	**58.65**	**68.32**	**47.42**	**62.92**	**69.90**	**55.19**
CLIP (w/ groundtruth box proposal)								
CPT-adapted [6]	24.16	24.70	25.07	22.28	28.68	25.12	23.39	28.42
GradCAM [5]	54.00	54.01	48.00	52.13	43.85	45.41	50.13	41.47
ReCLIP [6]	**65.48**	64.38	49.20	50.23	48.58	49.69	48.08	52.50
Ours	64.99	64.03	49.75	50.18	49.77	52.82	49.90	57.29
Ours+VR-CLIP	65.11	**66.00**	**58.65**	**64.78**	**53.98**	**65.60**	**68.59**	**63.51**
FLAVA								
Ours	60.95	59.99	48.89	50.02	46.86	49.37	47.76	51.68
Ours+VR-FLAVA	**61.25**	**60.86**	**50.79**	**53.35**	**47.62**	**52.46**	**52.66**	**52.92**
FLAVA (w/ box size prior)								
Ours	60.40	60.73	54.82	59.73	**48.25**	57.22	59.61	55.05
Ours+VR-FLAVA	**60.48**	**61.28**	**55.00**	**61.13**	48.17	**57.80**	**60.86**	**55.33**
FLAVA (w/ groundtruth box proposal)								
Ours	67.71	66.11	52.17	51.73	54.33	55.75	50.68	62.10
Ours+VR-FLAVA	**67.97**	**67.25**	**54.66**	**55.78**	**54.82**	**58.22**	**55.83**	**62.47**

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. Results excluding object boxes smaller than 5% of the image size are denoted as `w/ box size prior`. Results using groundtruth box proposals are indicated as `w/ groundtruth box proposal`. For every combination of model and box proposal type, the best results are highlighted in **bold**.

where no `object` is involved, which is a special case of our grounding pipeline.

ChatGPT plays an important role in improving the robustness of our grounding approach. In Fig. 2, as for "rt bottom chair", ChatGPT understands that "rt" stands for "right", allowing us to accurately generate triplets as depicted in the image. Similarly, in Fig. 3, it is worth highlighting the example "rider of the gray elephant". Here, ChatGPT made some reasonable deduction that rider is "on top of" the elephant. With longer captions, as shown in Figure 3, ChatGPT can consistently parse each entity, along with its complex attributes, affiliations, and inter-entity interactions, which are vital for accurate grounding. These examples demonstrate ChatGPT's superior robustness compared to ReCLIP's language parsing method, especially in challenging scenarios.

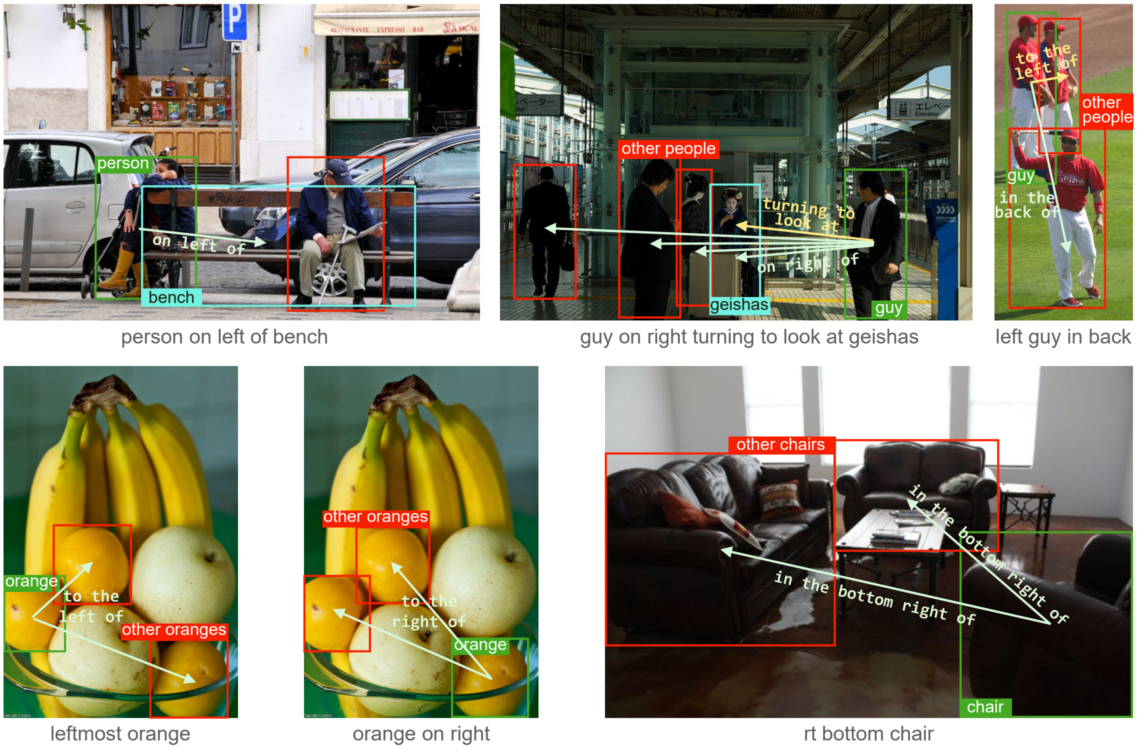

Figure 2. **Zero-shot visual grounding results on RefCOCO.** Our predictions are in green box, distraction objects are in red box. Arrows represent relationships between visual objects, and the text on the images are the parsed triplets.

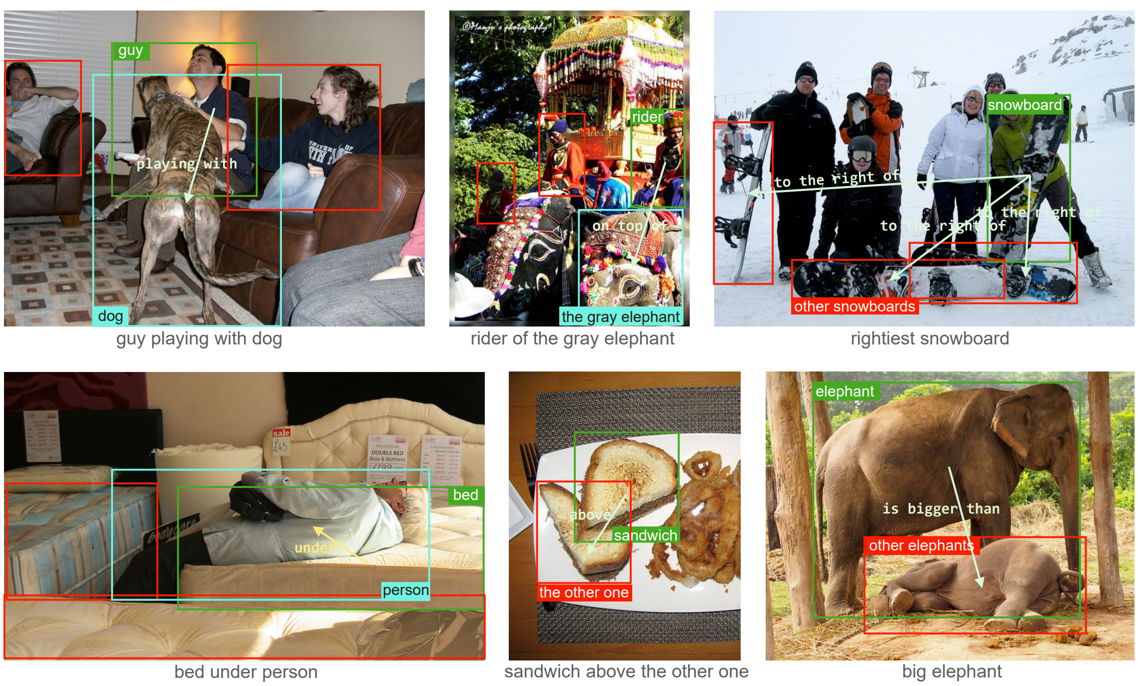

Figure 3. **Zero-shot visual grounding results on RefCOCO+.** Our predictions are in green box, distraction objects are in red box. Arrows represent relationships between visual objects, and the text on the images are the parsed triplets.

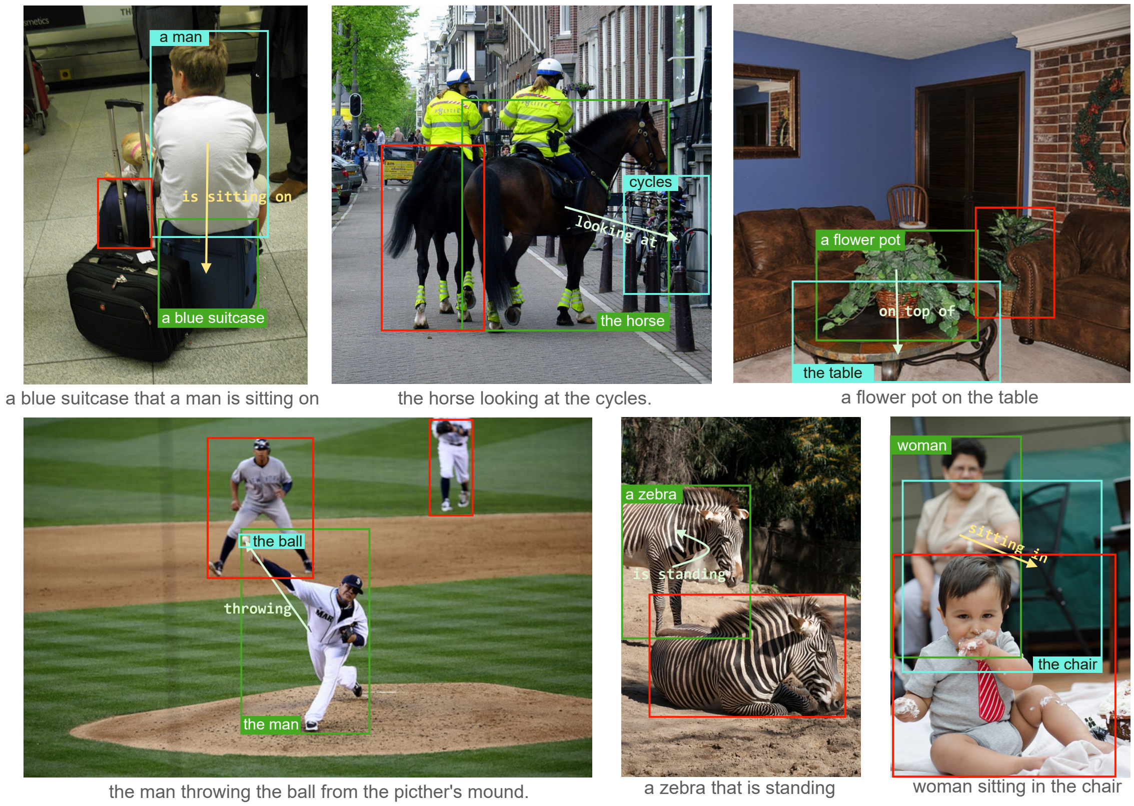

Figure 4. **Additional zero-shot visual grounding results on RefCOCOg.** Our predictions are in green box, distraction objects are in red box. Arrows represent relationships between visual objects, and the text on the images are the parsed triplets.

Figure 5. **Additional zero-shot visual grounding results on Who's Waldo.** Predicted annotation links are in the same color. Arrows represent relationships between visual objects, and the text on the images are the parsed triplets.

References

[1] 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. 1

[2] 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. 1

[3] 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

[4] 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. 1

[5] 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. 2

[6] 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. 2

[7] 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. 2