Learning Visual Grounding from Generative Vision and Language Model

Supplementary Material

A. Visualization and Analysis

Dataset. Figure A1 illustrates the object category distributions of generated VLM-VG dataset. We visualize some examples from the VLM-VG dataset with the three types of annotations in Figure A2. By human inspection, the incorrect or inaccurate annotations are labeled as red. We can see the regional captions generated by the VLM could generally provide detailed and accurate descriptions of the major object, *e.g*. in the third example in the second row, the VLM not only successfully recognizes the object as a computer monitor, but also captures the detail of the sticky notes on the monitor. However, the caption solely relied on cropped regions sometimes might miss or mistakenly describe the global scenes. *e.g*. in the third example in the first row, the caption successfully describes the major object as "a man in suit" but mistakenly recognizes the action as "standing in front of a window" since the cropped regions didn't contain the global information such as the ocean and ship. Besides, we also observe that although most of the relation annotations could provide correct spatial information to refer to the object, the simple rule-based method sometimes may still fail to generate the most appropriate spatial descriptions due to the complexity of the scene.

Besides the high quality of the automatically generated referring annotations, another major advantage of VLM-VG dataset is the diversity of the text annotations. By combing various types of annotations, VLM-VG can annotate one single object from multiple perspectives and at different levels of granularity, which matches human cognition and linguistic manners. Figure A3 shows three examples that contain multiple referring annotations varying from the level of detail to the angle of descriptions. Trained on the VLM-VG dataset with diverse annotations, the grounding model can achieve stronger robustness and generalizability.

Model prediction. We illustrate the model's zero-shot REC and RES predictions in Figure A4, A5, and A6. In detail, Figure A4 shows some examples on RefCOCO and RefCOCO+ datasets which use relatively short simple phrases as referring expressions. Trained on VLM-VG dataset, our model can successfully detect objects, understand spatial relations, and distinguish objects by their attributes accurately without seeing human-annotated grounding data. Figure A5 shows results on the RefCOCOg dataset which require models to understand longer and more complex sentences as referring expressions. The model also demonstrate a solid capability to associate objects with complex descriptions. For example, the second and third example in Figure A5 referring to two people in one image. The model successfully distinguished and located the two people with similar dressing yet different actions, indicating the model's fine-grained reasoning capability.

In order to better understand the shortcomings of the model, we collect several representative failed examples as illustrated in Figure A6. One of the major failure modes is that the model fails to capture the visual details mentioned in the referring expression, *e.g*. the "white table" in the corner in the first image and "strawberries" in the last images. Moreover, we also observe that complex scenes, such as the one depicted in the third image, pose challenges for the model to locate the correct object by spatial relationships. Furthermore, the third example in the second row reveals a potential limitation of the VLM-VG dataset: it may not cover all the intricate relationships present in real-world scenarios which are hard to be captured by the simple rule-based relation modeling method.

Caption pipeline comparison. A concurrent work [11] also utilized object bounding boxes to generate high-quality image descriptions. It employed VLMs to generate captions for the entire image, using object bounding boxes for image parsing and fact-checking, followed by refining the captions for more detailed and accurate descriptions. While this checking and refining pipeline offers more fine-grained visual details compared to our cropped captions, our method strikes a balance between quality and scalability, focusing on modeling object relationships and identifying distinguishable attributes that facilitate object identification and reference.

B. Limitations

When generating the referring expressions, we utilize the rule-based methods utilizing localization heuristics. The manually designed rules as rough approximations for relation on three dimensions empirically show a huge improvement on grounding models' spatial awareness. However, the simple rule-based relation modeling may fail under the complex scenario. For example, when there are adjacent objects with same category, *e.g*. person, the method may produce annotations such as "person to the left of person" which is not distinctive enough to refer to a specific object. Besides, simply comparing center coordinates and box size to model horizon and depth relation might cause incorrectness due to ignoring the intrinsic size of different objects and struggle with more complex and diverse spatial relations.

Additionally, we scale grounding datasets based on detection datasets which are generally one to three orders of

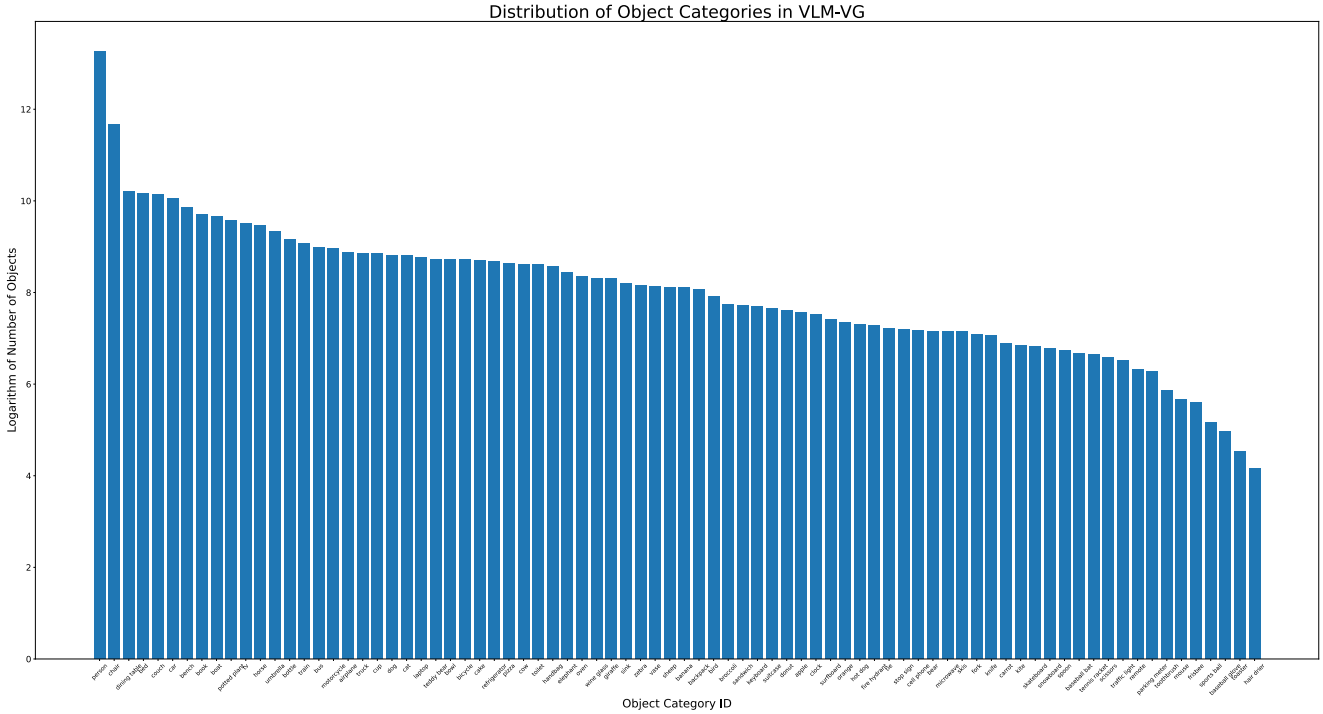

Figure A1. **Object category distribution in the VLM-VG dataset.** The dataset consists of 512K images, 1.1M objects, and 16.2M text annotations. The category distribution approximately follows a Zipfian distribution, consistent with natural image datasets observed in prior works [12].

magnitude larger, without relying on expensive and inflexible manual text annotations. This may be a limitation in the long-run when we want to scale up visual grounding models to objects beyond what are available in current detection datasets. We conducted some initial exploration in Table 5 and observe promising scaling behavior.

C. Additional Implementation Details

Relation modeling. In Section 3.2, after generating the relation tuple `(noun, rel, noun)` and `(noun, rel)` for relative and absolute relationship respectively, we use pre-defined templates to formulate the phrases based on the tuple. The templates are listed in Table A1.

Dimension	Tuple	Templates
Horizontal	`(A, left, B)`	*A to the left of B*
	`(A, right, B)`	*A to the right of B*
	`(A, left)`	*A left / left A*
	`(A, right)`	*A right / right A*
	`(A, left most)`	*A on the far left / A far left / far left A*
	`(A, right most)`	*A on the far right / A far right / far right A*
	`(A, middle)`	*A middle / middle A / center A / A center*
Vertical	`(A, top)`	*A top / top A*
	`(A, bottom)`	*A bottom / bottom A*
Depth	`(A, behind)`	*A behind / behind A*
	`(A, front)`	*A front / front A*

Table A1. Templates to formulate spatial relation phrases.

Attribute	Prompt
cloth	*What is the person wearing?*
gender	*What is the person's gender?*
identity	*What is the identity of the person?*
action	*What is the {class} doing?*
color	*What is the color of the {class}?*
material	*What is the material of the {class}?*
shape	*What is the shape of the {class}?*

Table A2. Prompts to query PaLI-3 for each attribute. {*class*} denotes category name.

Attributes modeling. When generating the attribute-rich annotations, we choose 7 types of attributes and query PaLI-3 with the corresponding prompts as detailed in Table A2. For each attribute, not all the object categories are applicable to the attribute. In details, for the 80 COCO classes, *["cloth", "gender", "identity"]* are applicable to the class *human*, *"action"* is applicable to the class *[person, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe]*, *"material"* is applicable to the class *[bench, backpack, umbrella, handbag, tie, suitcase, sports ball, bottle, wine glass, cup, fork, knife, spoon, bowl, chair, couch, bed, dinning table, toilet, sink, clock, boat, vase]*, *"shape"* is applicable to the class *[stop sign, parking meter, bench, handbag, suit-*

case, kite, bottle, cup, bowl, dining table, couch, bed, toilet, clock, vase], and *"color"* is applicable to all the classes. We only query the VLM to model the applicable attributes for each object.

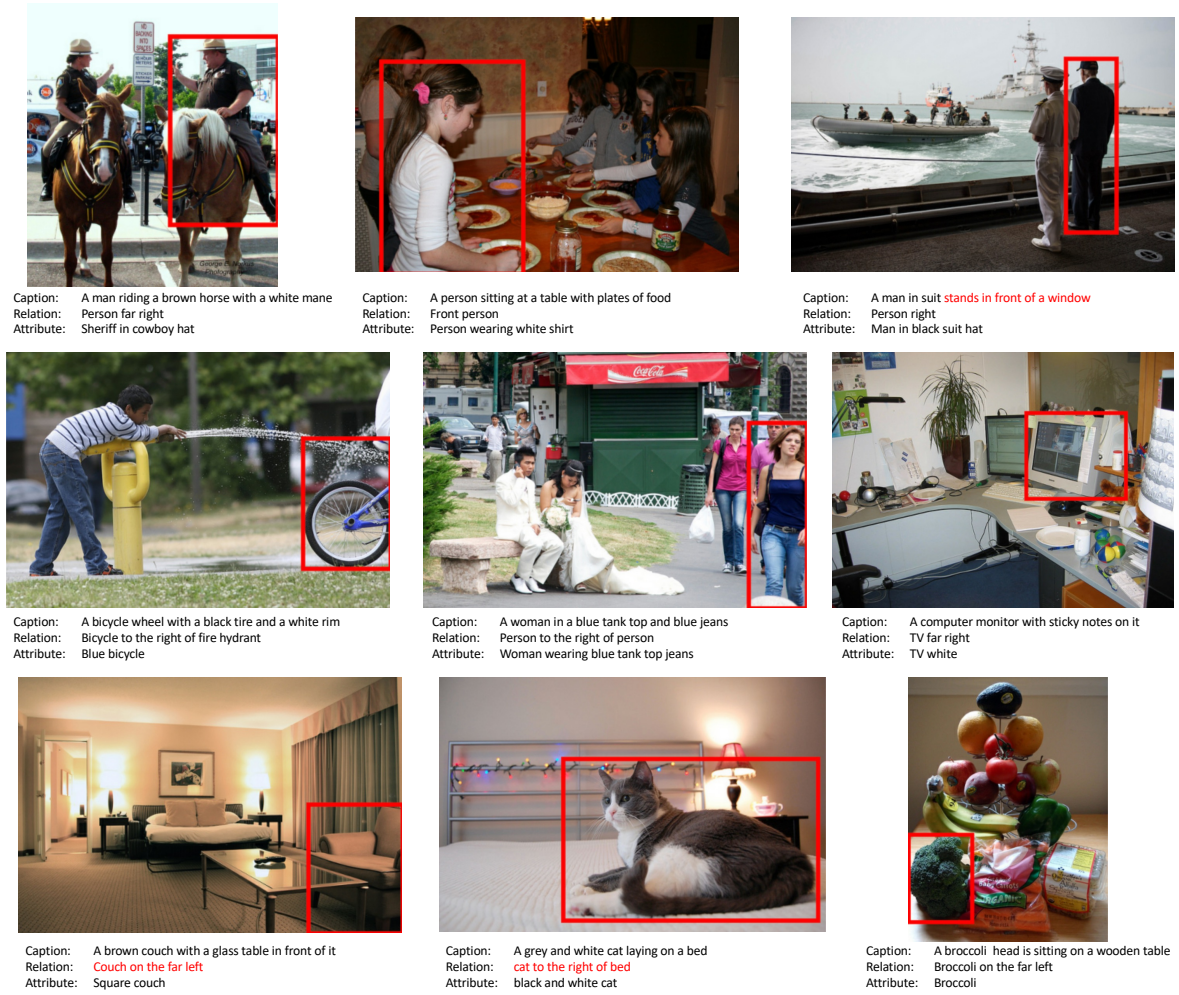

Figure A2. **Visualization of VLM-VG dataset.** By human examination, the incorrect or inaccurate annotations are colored red.

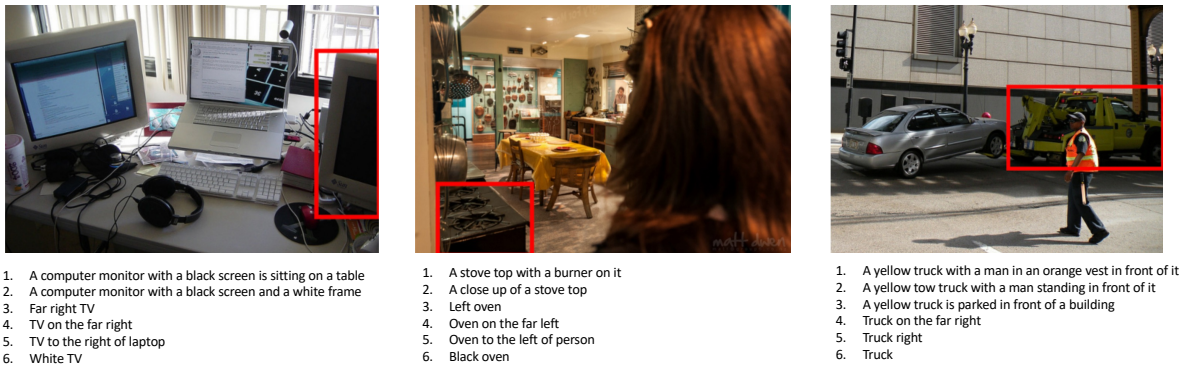

Figure A3. **Diversity of generated annotations.** Our VLM-VG dataset provides referring expressions annotations from multiple perspectives aligning with human linguistic manners.

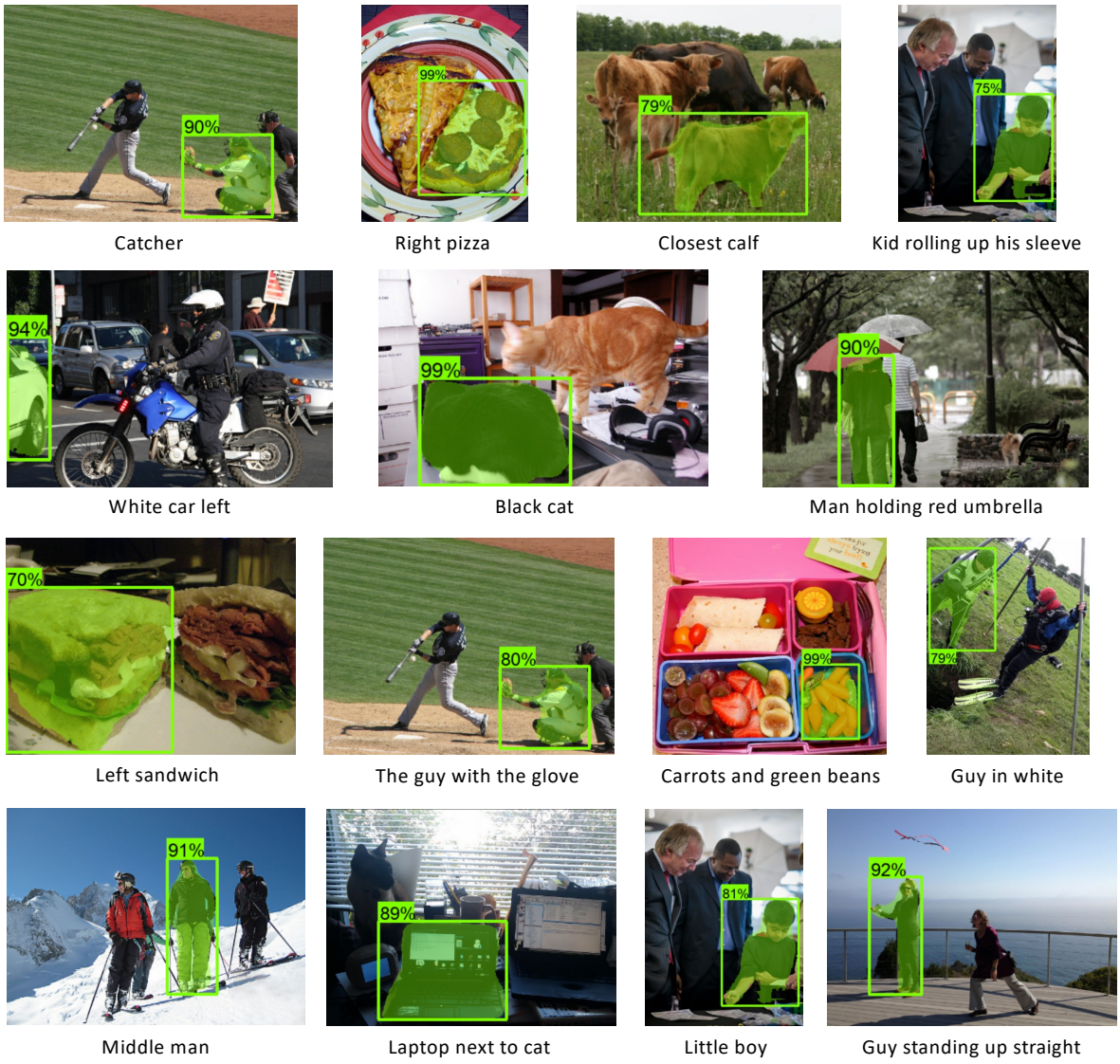

Figure A4. **Visualization of the zero-sot REC and RES predictions on RefCOCO and RefCOCO+.** RefCOCO dataset requires spatial relationship understanding.

Figure A5. **Visualization of the zero-sot REC and RES predictions on RefCOCOg.** RefCOCOg requires models to understand longer and more complex referring expressions.

Figure A6. **Failure cases of the model prediction.**