GeoChat: Grounded Large Vision-Language Model for Remote Sensing

Supplementary Material

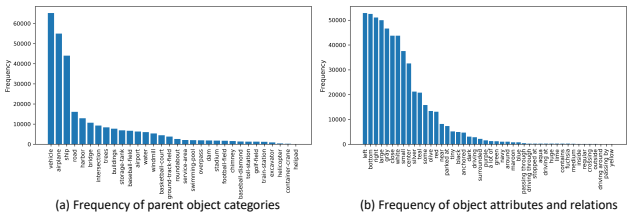

Figure 1. Frequency distribution for object attributes and classes.

Model	Dataset	BLEU-4	ROUGE-L	METEOR
RSGPT	UCM-Captions	65.7	78.3	42.4
GeoChat		**80.7**	**79.1**	**75.4**
RSGPT	RSICD	36.8	53.3	30.3
GeoChat		**57.4**	**56.3**	**49.3**

Table 1. Image Captioning: UCM captions and RSICD dataset, after five epochs.

Model	Presence	Comparison	Rural-Urban
RSGPT	91.17	91.70	94.00
GeoChat	**91.81**	**92.15**	**95.00**

Table 2. Results on LRBEN after finetuning for two epochs.

1. Dataset Distribution

Fig. 1 (a) shows data diversity via the frequency plot of each parent object category in our dataset. It exhibits a long-tail distribution, with vehicles, airplanes, and ships being frequent classes, whereas helicopters, cranes, and helipads are rare. Fig. 1 (b) shows the frequency of each attribute and relationship in our dataset. The color and size attributes occur the most. Further analysis will be added in the camera-ready.

2. Quality of pseudo labels

We first manually annotated 100 samples of our dataset for LoveDA classes. These annotated samples served as a benchmark for refining the pseudo labels, employing a systematic five-step approach. After each step, three people assess the quality of these pseudo labels. E.g., we skipped any images containing categories where predictions of ViTAE-RVSA [31] are consistently poor. To lower false positives further, we excluded predictions for which ground truth annotations were already available in our dataset. Subsequently, we considered only those predictions whose areas fell within the 20% to 80% percentile range of their respective class areas, as determined by manual annotations. Lastly, we incorporated class-specific checks by scrutinizing predictions, excluding building predictions that fell within valid water or road predictions. To further reduce the impact of noise, we only compute pseudo labels for 50% of our dataset.

3. Supervised Finetuning

Further, we finetuned GeoChat for three supervised tasks, i.e., scene classification, image captioning, and VQA. We follow RSGPT [13], and finetune on UCM-Captions and RSICD for five epochs (Table 1) and on LRBEN for two epochs (Table 2). Further, we get 89.13% accuracy on the AID dataset and 94.71% on the UCMerced dataset after just one epoch of finetuning. Our approach consistently achieves SoTA results. Table 1 shows the supervised finetuning results on image captioning tasks.

4. Additional Qualitative Examples

In this section, we delve further into the qualitative examples and capabilities of GeoChat. Figure 2 not only demonstrates the proficiency of GeoChat in providing detailed image descriptions and precise bounding boxes but also emphasizes a notable feature — the prediction of rotated bounding boxes, which sets it apart from Minigpt-v2[1]. GeoChat provides both the image description as well as the bounding boxes for the objects detected. Moving on, Figure 3 presents some qualitative examples on region-based captioning. When given a bounding box and the instruction prompt '[identify]', GeoChat is able to provide concise brief description about the regions. Figures 4 and 5 present referring expression grounding results, emphasizing the model's ability to accurately understand and respond to specific references within images. Figures 6 and 7 show qualitative examples for the instruction-following abilities of GeoChat, as well as the ability to handle multiple questions in a single conversation.

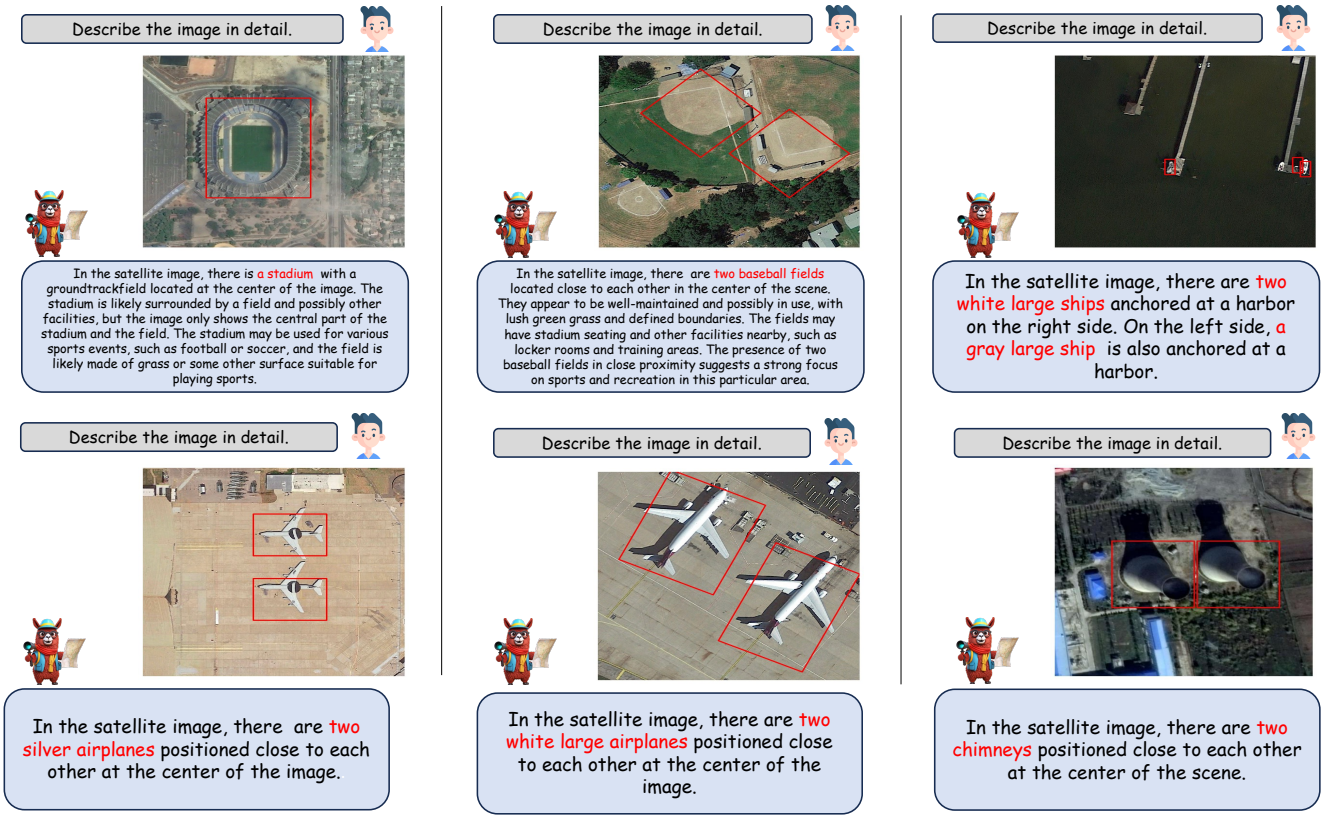

Figure 2. Qualitative examples of grounded image captions. GeoChat outputs both the description of the image as well as the bounding boxes for all the objects detected. These images are a part of the proposed evaluation dataset for grounded captioning.

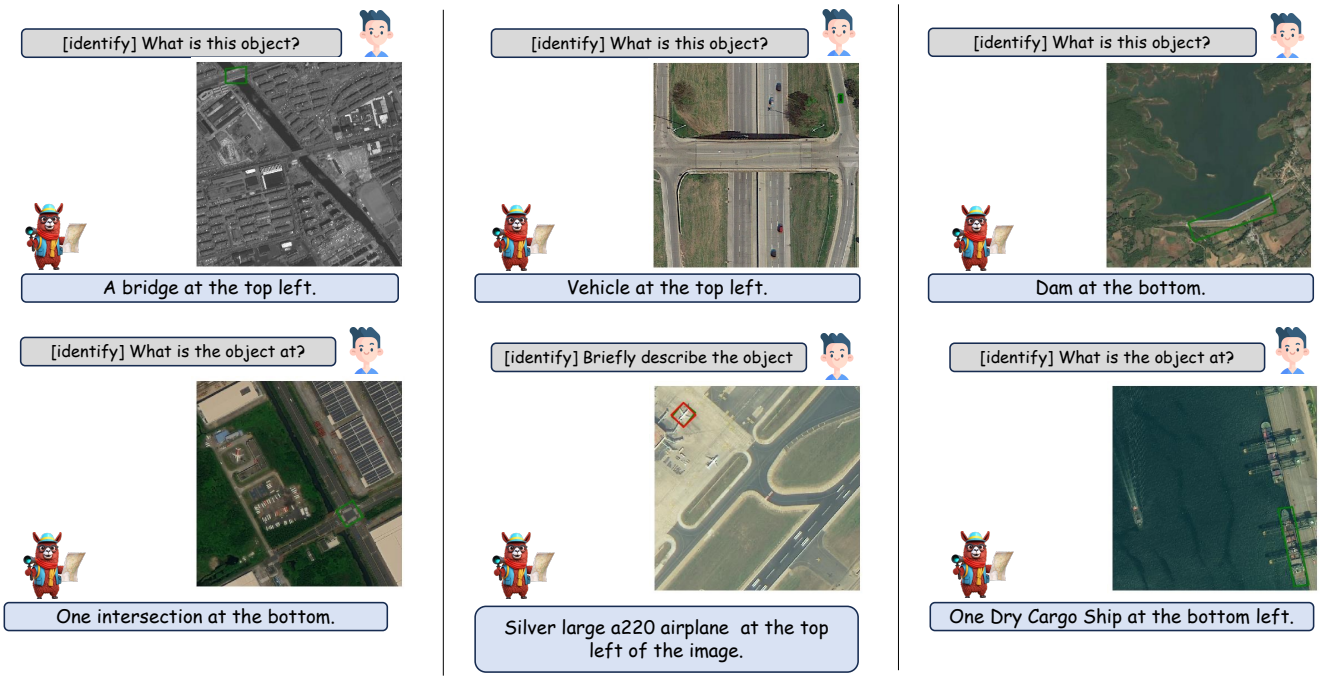

Figure 3. Qualitative examples for region-based captioning. Given a bounding box, GeoChat is able to provide brief descriptions about the area or the object covered by the bounding box. These images are a part of the proposed evaluation dataset for region based captioning.

Figure 4. Qualitative examples of referring expression grounding. Given an object, GeoChat is able to locate objects and draw rotated bounding boxes correspondingly. These images are a part of the proposed evaluation dataset for referring expression grounding.

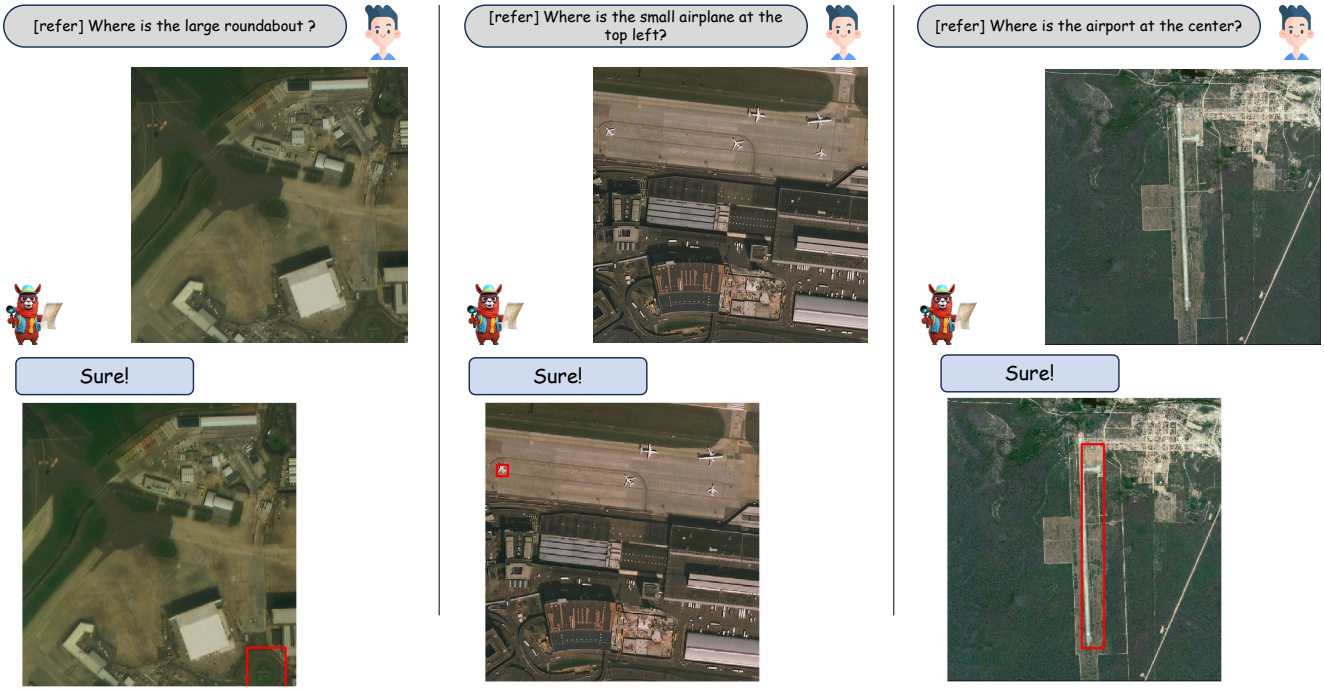

Figure 5. More qualitative examples of referring expression grounding.

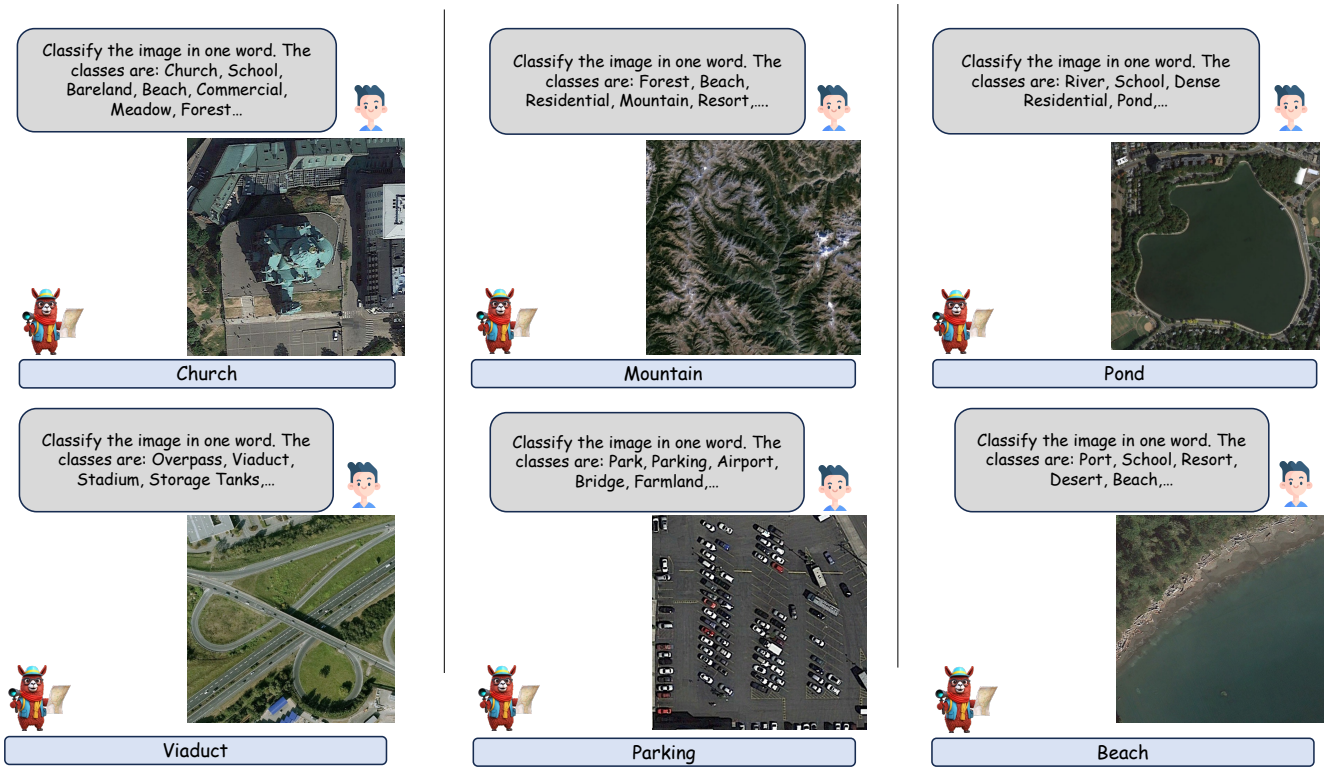

Figure 6. Qualitative examples for scene classification. We give the model all the classes from the dataset and ask to choose only one. These results are from AID[3] dataset, with 30 classes.

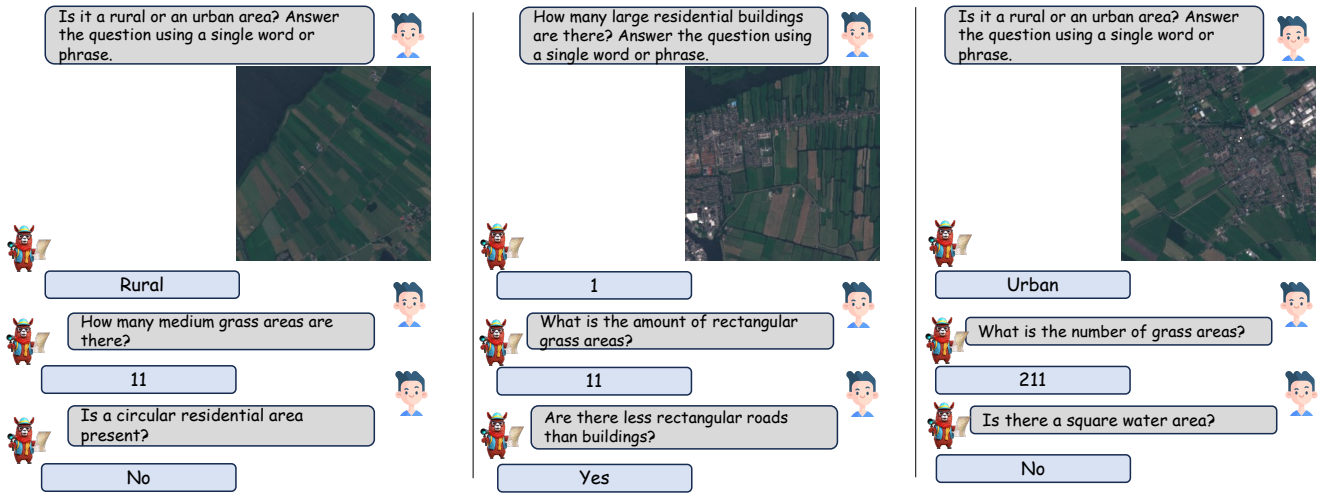

Figure 7. Qualitative examples for Visual Question Answering tasks. GeoChat is able to hold multi-turn conversations, based on various types of questions, including presence, count, complex comparisons and so on. It performs well on low-resolution images as well. These examples are from the test set of RSVQA-LRBEN[2] dataset.

References

[1] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1

[2] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12):8555–8566, 2020. 4

[3] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 4