Overview

In this supplementary material, we provide following items:

- (Sec.1) Details about the preliminary experiment.
- (Sec.2) More results of self-consistency enhancement.
- (Sec.3) Visualization of the *SC-Tune* effects.
- (Sec.4) Implementation details of training data selection.

1. Preliminary Experiment Details

In this section, we delineate the details of our preliminary experiment. Initially, we randomly select 4K bounding boxes (bbox) from three datasets (*i.e*. RefCOCO [7], OpenImage [3] and Object365 [5]) to respectively build the test splits, ensuring that each bbox originates from different images. For any given baseline model, we fill the image along with the coordinates of the bbox into its referring expression generation (REG) instruction template, guiding the model to generate a region caption for the specified bbox. Subsequently, we fill the image and the generated caption into its referring expression comprehension (REC) instruction template, enabling it to locate the coordinates described by the caption. Following previous works [4, 6], we compute the intersection over union (IoU) between newly predicted coordinates and the original ones. If the IoU exceeds 0.5, the model is considered to have achieved the required level of self-consistency for the given sample. Ultimately, we employ the proportion of samples meeting this criterion within the test split (also known as $Pr@0.5$) as the metric for measuring the self-consistency level of a model.

2. Self-Consistency Enhancement

In this section, we present the full results across baseline models [1, 2] and data distributions [3, 5, 7] to evaluate the self-consistency enhancement brought by *SC-Tune*. For the two baseline models, both OpenImage and Object365 serve as out-of-domain data sources, whereas RefCOCO remains visible throughout their training process. We conduct data filtering (detailed in Sec 4) on OpenImage and Object365, aligning their data sizes to be comparable with that of RefCOCO for fairness. It is worth to emphasize that prior to the filtering process, we have already excluded the 4k test samples utilized in the preliminary experiment. Ultimately, we obtain about 166K training samples from Object365 and 138K training samples from OpenImage, respectively. These samples, in conjunction with the training split of RefCOCO, are utilized for the self-consistency enhancement evaluation shown in Table 1.

Table 1. Self-consistency evaluation on three benchmarks. We report self-consistency level using accuracy, where a sample is considered as right when IoU between prediction and ground-truth is higher than 0.5.

Method	Object365	OpenImages	RefCOCO
Qwen-VL	76.9	52.9	87.9
+*SC-Tune* (Object365)	**94.1**	68.8	**93.8**
+*SC-Tune* (OpenImages)	89.6	**73.6**	92.0
+*SC-Tune* (RefCOCO)	83.4	58.8	93.5
MiniGPT-v2	65.4	38.2	87.2
+*SC-Tune* (Object365)	**76.6**	48.5	90.4
+*SC-Tune* (OpenImages)	74.9	**50.2**	91.1
+*SC-Tune* (RefCOCO)	69.8	45.5	**91.6**

It clearly indicates that employing *SC-Tune* enhances the self-consistency levels for both baseline models. Taking Qwen-VL [1] as an example, tuning with the out-of-domain datasets (*i.e*. Object365 and OpenImage) results in a self-consistency level increase of over 18% in respective test splits. Moreover, applying *SC-Tune* on in-domain data RefCOCO similarly elevates self-consistency levels across three test splits. The results sufficiently demonstrate the robustness and applicability of *SC-Tune* across different data sources. Additionally, a similar pattern is observable in MiniGPT-v2, further confirming the compatibility of *SC-Tune* to various models.

3. Visualization

In this section, we leverage Qwen-VL as an example to demonstrate the improvements of REG and REC capabilities after the application of **SC-Tune**. Figure 1 qualitatively showcases the enhanced REG capability, which can be summarized in two aspects:

(1) A refined understanding of detailed and unique attributes. Specifically, as illustrated in the upper part of Figure 1, the original model generates a generic description, which is applicable to all three women depicted in the given image without distinction. However, after *SC-Tune*, the model captures more detailed information about the selected object, such as age, body parts, and facial orientation for unique identification.

(2) An enhanced capability to integrate visual context. For instance, in the lower part of Figure 1, compared with the description generated by original model, the *SC-Tuned* model incorporates additional contextual information. It includes sufficient contextual clues like the jersey number of the player and his interaction with surrounding individuals.

Figure 2 qualitatively showcases the enhanced REC capability. The cases illustrated in the figure demonstrates that for a given bbox and image, even if two models, before and after *SC-Tune*, generate informative and equal descriptions, the model without *SC-Tune* still fails to accurately locate back the original bbox. It is largely attributes to the

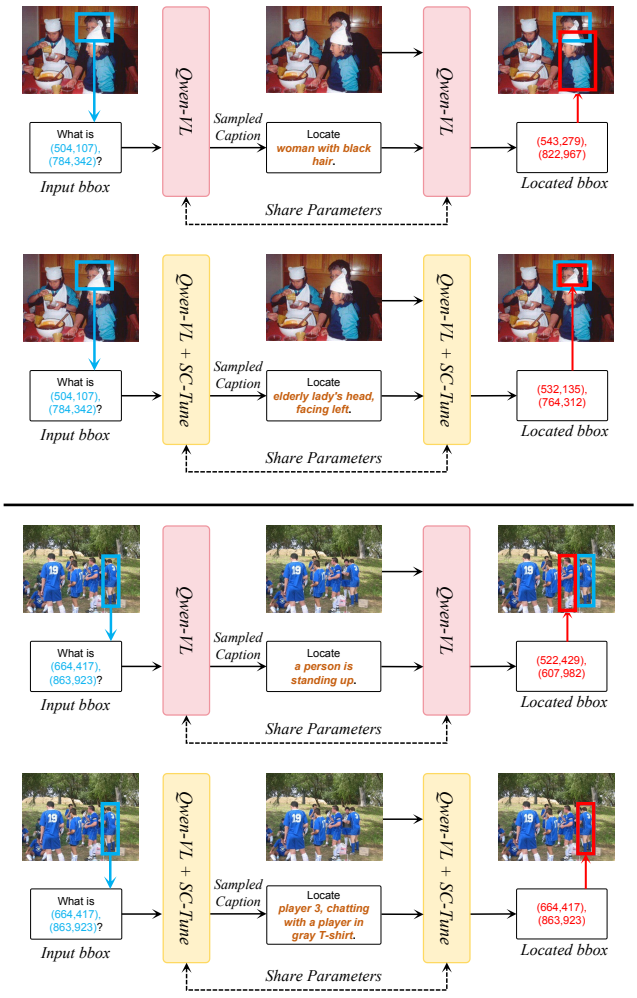

Figure 1. Case study to demonstrate the enhanced REG capability after *SC-Tune*.

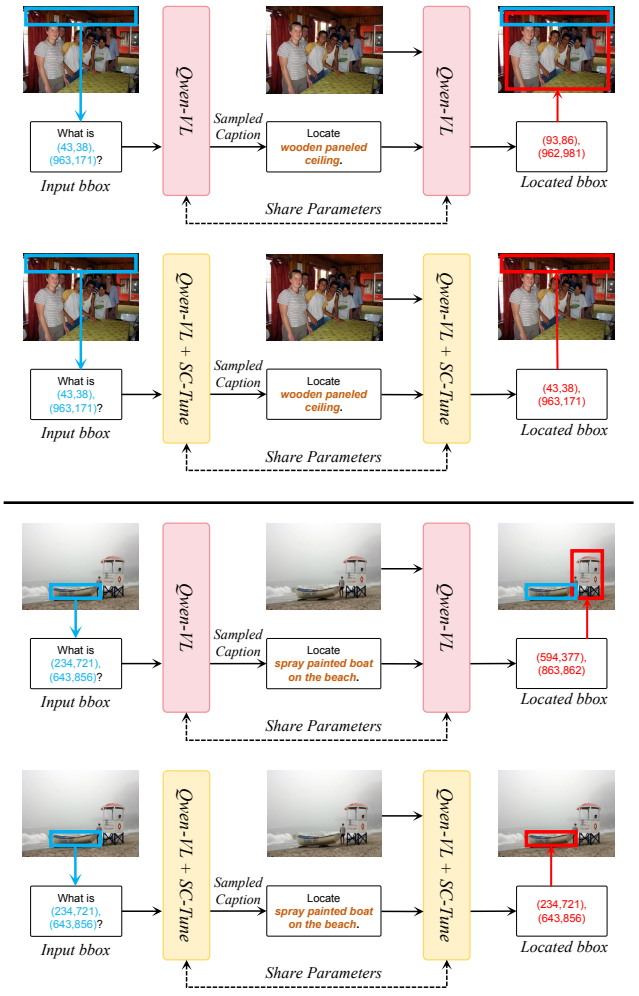

Figure 2. Case study to demonstrate the enhanced REC capability after *SC-Tune*.

lack of self-consistency in the multi-task learning paradigm, which hinders the effective alignment of these two capabilities *i.e*. REC and REG. *SC-Tune* significantly amends this mismatch, thereby enhancing the visual grounding ability.

4. Training Data Filtering

In this section, we introduce the process of data filtering. Specifically, for each image, if there are no categories that appear more than once, the image and all the corresponding annotations are discarded. Otherwise, we record the information as a triplet consisting of the image, category, and all bboxes that fit the category. This design is intended to better introduce ambiguity and increase the difficulty of training. Additionally, for the bboxes within an image, if two bboxes have an intersection over union (IoU) greater than 0.5, the triplets associated with these overlapping bboxes are removed. It is to address the potential confusion caused by overlap, which could interfere with model training. Lastly, we eliminate the triplet which contain a bbox that occupy less than 2% of the image area. The same filtering approach is applied to both the Object365 and OpenImage datasets.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[2] Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 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

[3] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 1

[4] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306, 2023. 1

[5] 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. 1

[6] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15325–15336, 2023. 1

[7] 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. 1