Supplementary Materials for Towards Open-Ended Visual Recognition with Large Language Models

Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen

ByteDance

In the supplementary materials, we provide more technical details of OSM. We also include more quantitative results and qualitative results, along with comparisons with GPT-4V [9]. Moreover, we show that OSM can also be easily extended with part-level and box-level datasets, further unleasing the potential of OSM.

Instruction Template We summarize the instruction template we used for OSM training in Tab. 1.

Dilemma between Accuracy and Generalization We also train OSM under different seen number of masks (*i.e.*, 1, 3, 6, 9 millions respectively), as shown in Fig. 1. Empirically, we consider Acc as a metric to measure how well the model can accurately recognize the object and NIV as a metric to measure the generalization ability. We note that there exists a dilemma between the accuracy and generalization, *i.e.*, when the number of seen masks increases, we notice that the model achieves higher accuracy while inevitably having a higher overfitting to the training vocabulary, and predicts in a more conservative manner. From 6M to 9M, the accuracy improvement majorly comes from the decrease of NIV. We note that how to ensure an accurate object recognition while avoid overfitting to the training vocabulary is an interesting future research problem.

Incorporating Part- and Box-level Datasets It is worth noting that OSM seamlessly accommodates part-level and box-level datasets, further enhancing its versatility. To enhance OSM for part-level and box-level recognition (note that OSM already shows emergent part recognition ability as illustrated in Fig.1 of main paper, but we believe introducing such datasets could further advance its ability), we introduce PartImageNet [4], Pascal-Part [2], and V3Det [12] datasets into the training data. For part data, we prepend the object name to part name, in case many parts sharing the same names (*e.g.*, in PartImageNet, many different classes may have the same part named *head*). We also remove those class names which are too vague (*e.g.*, *train left side*, *bus upper side* in Pascal-Part). For detection data, we consider the bounding-box as a box-shaped binary mask and thus is easily unified into OSM. Additionally, we augment the panoptic/instance segmentation data (*e.g.*, COCO, LVIS) by randomly converting each segmentation mask into its corresponding bounding box. In cases where a bounding box serves as input, we appropriately adjust the instruction by replacing the term “segmentation mask" with “bounding box." It's worth mentioning that we do not include image-level data (*e.g.*, ImageNet) at this stage, as the semantic label could introduce bias when there exist multiple

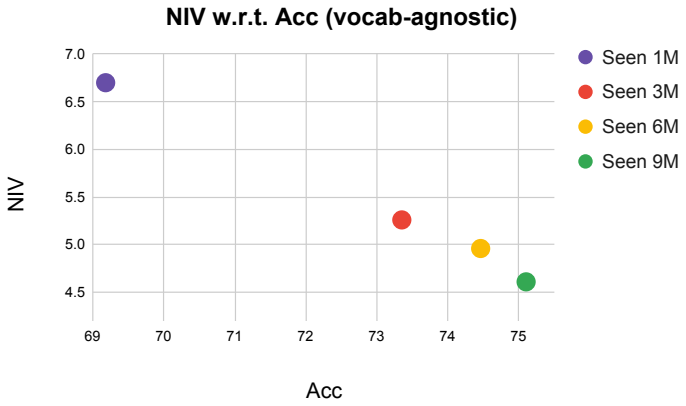

Fig. 1: NIV w.r.t. Acc, when number of seen masks varies. The Acc increases as the number of seen masks increases, showing that the model is better trained to fit the target dataset. However, its NIV score becomes much lower, indicating that the model is losing the generalization ability.

objects yet sharing single label. We demonstrate their effects in Fig. 2, where we use SAM [5] and DETA [10] as the proposal model respectively.

Qualitative Results We explore the application of OSM on top of SAM [5] and kMaX-DeepLab [15], and provide qualitative results, which are presented in Fig. 3 and Fig. 4 respectively. These results underscore the superiority of OSM in practical scenarios and its potential to demonstrate open-ended recognition with fine-grained masks. When obtaining mask proposals from SAM [5], we use the SAM variant with ViT-H [3] backbone, with points per side 32, IoU threshold 0.95, stability threshold 0.95, and minimum mask size 800. This helps avoid too many small masks that are not recognizable (*e.g.*, super-pixel level masks). When obtaining mask proposals from kMaX-DeepLab [15], we use the one trained on COCO Panoptic dataset with ConvNeXt-L [7] backbone, and we set the "thing" and "stuff" threshold to 0.1 to obtain more mask proposals and feed them into OSM. Afterwards, we apply mask-wise post-processing following [14, 15].

Comparison against GPT-4V We provide a qualitative comparison between GPT-4V and OSM. We follow [13] to prompt GPT-4V for mask recognition. Specifically, we highlight the mask boundaries as auxiliary cues in the image, and annotate each mask center with a numeric ID. We feed the prompted image to GPT-4V, along with text prompt "*I have labeled a bright numeric ID at the center for each visual object in the image. Please enumerate their names* (i.e.*, semantic class) with one, two, or three words.*". The results are shown in Fig. 5, with first column showing the image after mask prompting and fed to GPT-4V, and second column for GPT-4V predictions, third column for OSM predictions. We observe that OSM has a more accurate prediction compared to GPT-4V (*e.g.*, in the first row, OSM correctly predicts mask 5 and 11 as *bench* and *fence*, while GPT-4V wrongly predicts them both as *streetlight*), which is

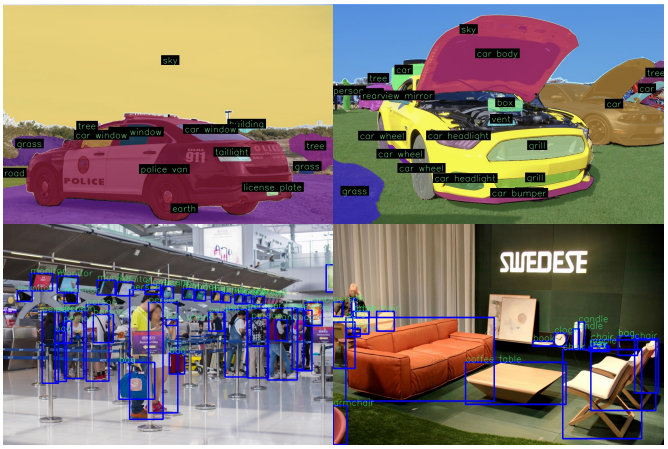

Fig. 2: Extending OSM with part-level and box-level dataset. We note that the OSM framework is general and we can easily extend it with part-level and box-level data, leading to a stronger performance and more diverse usage. Best viewed zoom in to see predicted class names.

often confused by the context (*e.g.*, in the first row, for the mask 10, GPT-4V predicts *buildings* instead of *mountain*, potentially due to confusion from the buildings below). However, we also note that OSM's prediction is more conservative compared to GPT-4V, which can predict a more specific word. For example, in the second row, GPT-4V predicts *man in armor* for the armed man in the image while OSM still predicts in a safer way with *person*. This also suggests a potential improvement of OSM from a stronger base model (*e.g.*, Llama2 [11]) or larger datasets with a better trade-off between accuracy and diversity [8]. We also apply similar strategy to test with state-of-the-art open-sourced multi-modal large language model (*e.g.*, LLava-1.5 [6], MiniGPT-v2 [1]) yet find them fail at generating reasonable outputs, coinciding observations in [13].

Visualization of NIV Cases To better understand what are NIV (Not-in-Vocab) cases, we visualize them using ground-truth mask against ground-truth annotation for COCO *val* set and ADE20K *val* set in Fig. 6 and Fig. 7 respectively. We note that with a pre-defined vocabulary, even the ground-truth annotations are usually biased and limited, where annotators have to pick a most similar class in the given vocabulary (*e.g.*, all *monitor* are labeled as *tv* in COCO). The biases could be learnt and inherited in existing closed-vocabulary and open-vocabulary models. However, we observe the OSM can predict a more appropriate class name without limitation of a given vocabulary, demonstrating the necessity and effectiveness of getting rid of a pre-defined vocabulary and pursuing open-ended visual recognition.

Table 1: Instruction templates. We randomly select one instruction template and insert the ground-truth class name during training. Only the first template *What is in the segmentation mask?* is used during testing.

1. What is in the segmentation mask? Assistant: {class_name}
2. Describe what is in the segmentation mask. Assistant: {class_name}
3. What does this segmentation mask show? Assistant: {class_name}
4. What is this segmentation mask? Assistant: {class_name}
5. What is the segmentation mask region of the image? Assistant: {class_name}
6. Briefly describe what you perceive in the segmentation mask region. Assistant: {class_name}
7. Please tell the category of what is indicated by the segmentation mask. Assistant: {class_name}
8. What does this segmentation mask segments? Assistant: {class_name}
9. What does this segmentation mask capture? Assistant: {class_name}
10. Answer the name of what is in the segmentation mask region. Assistant: {class_name}
11. What is the semantic class of the area given the segmentation mask? Assistant: {class_name}
12. Can you describe what is in the segmentation mask region? Assistant: {class_name}
13. From the image and segmentation mask provided, tell the category of the indicated region. Assistant: {class_name}
14. Could you use a few words to describe what is in the segmentation mask region? Assistant: {class_name}
15. Given the image and segmentation mask, answer what is in the region. Assistant: {class_name}
16. Tell me what you see in the segmentation mask region. Assistant: {class_name}
17. What can you see in the segmentation mask region? Assistant: {class_name}
18. Let me know what you can perceive in the mask region. Assistant: {class_name}
19. Give me the name of the object in the segmentation mask. Assistant: {class_name}

Fig. 3: Qualitative results on SA-1B dataset [5] of OSM, using SAM as the mask proposal model. Best viewed zoom in to see predicted class names.

Fig. 4: Qualitative results on SA-1B dataset [5] of OSM, using kMaX-DeepLab as the mask proposal model. Best viewed zoom in to see predicted class names.

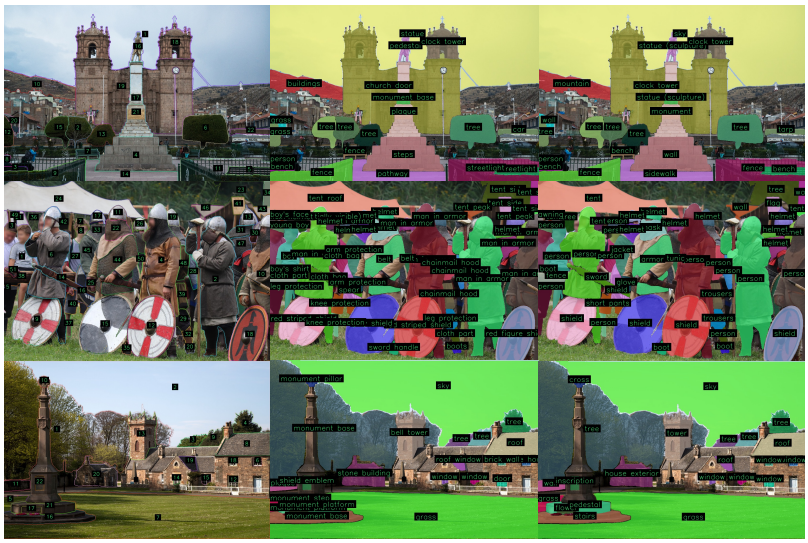

Fig. 5: Qualitative comparison *vs.* GPT-4V. We follow [13] to prompt GPT-4V for mask classification. Note that the enhanced image w/ mask prompting is only for GPT-4V input while OSM still takes the original image as input. Best viewed zoom in to see predicted class names.

Fig. 6: Visualization of NIV cases on COCO *val* set using ground-truth mask. We note that OSM can predict a more felicitous class compared to ground-truth, where annotations are limited to the fixed vocabulary and thus usually less expressive. We highlight the mask region with bounding box for better visualization purposes. Best viewed zoom in.

Fig. 7: Visualization of NIV cases on ADE20K *val* set using ground-truth mask. We note that OSM can predict a more felicitous class compared to ground-truth, where annotations are limited to the fixed vocabulary and thus usually less expressive. We highlight the mask region with bounding box for better visualization purposes. Best viewed zoom in.

References

1. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
2. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR (2014)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
4. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: ECCV (2022)
5. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV (2023)
6. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
7. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR (2022)
8. Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L.: Improving multimodal datasets with image captioning. arXiv preprint arXiv:2307.10350 (2023)
9. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
10. Ouyang-Zhang, J., Cho, J.H., Zhou, X., Krähenbühl, P.: Nms strikes back. arXiv preprint arXiv:2212.06137 (2022)
11. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
12. Wang, J., Zhang, P., Chu, T., Cao, Y., Zhou, Y., Wu, T., Wang, B., He, C., Lin, D.: V3det: Vast vocabulary visual detection dataset. In: ICCV (2023)
13. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
14. Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In: CVPR (2022)
15. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means Mask Transformer. In: ECCV (2022)