# InstaGen: Enhancing Object Detection by Training on Synthetic Dataset

## Supplementary Material

In this supplementary document, we present the experimental results of the LVIS-OVD benchmark in Section S1. Additionally, we perform an ablation study to evaluate the coupling between the diffusion model and the grounding head in Section S2. Furthermore, we evaluate the quality of the pseudo-labels generated by the grounding head in Section S3. Lastly, we provide more qualitative results in Section S4.

## S1. Open-vocabulary setting on LVIS

**Experimental setup.** We conduct experiments on the LVIS-OVD benchmark. The latest LVIS v1.0 [4] consists of 1203 categories, each with bounding box and instance mask annotations. The categories are divided into three groups based on the number of images in which each category appears in the training set: rare (1-10 images), common (11-100 images), and frequent (more than 100 images). In line with the problem setting in ViLD [3] and Detic [7], we treat the frequent and common classes as base categories, while considering the rare classes as novel categories. For evaluation on LVIS v1.0 *minival* set, we mainly consider the mask Average Precision for novel categories, *i.e.* $AP_{novel}$. However, to complete the AP metric, we also report $AP_c$ (for common classes), $AP_f$ (for frequent classes) and AP (for all classes).

Similar to PromptDet [2], we enhance the prompt template by incorporating a more detailed description to mitigate lexical ambiguity, particularly for the rare classes in LVIS. It should be noted that the description can be easily extracted from the metadata of the dataset. Consequently, the text prompt for the selected categories is generated as follows: 'a photograph of [category1 name] ([category1 description]) and [category2 name] ([category2 description])'. During the training of the grounding head, we utilize 500 synthetic images per category per training epoch. In addition, for the training of the object detector, we employ 250 synthetic images per category per training epoch and conduct 24 epochs of training.

**Comparison to SOTA.** We conduct a comparison with the existing CLIP-based open-vocabulary object detectors using the Mask-RCNN model with ResNet-50, as shown in Table S2. The results indicate that our detector, trained on synthetic dataset from **InstaGen**, achieves comparable or improved performance over existing CLIP-based methods.

## S2. Tight coupling *vs*. Loose coupling

To generate high-quality bounding-boxes for the synthetic images, we have designed a tight coupling between the diffusion model and the instance-level grounding head, namely, the grounding head predicts the bounding-boxes based on the SDM's internal representation. To demonstrate the effectiveness of the tight coupling design, we compare it with a loose coupling design. For the latter, we train an open-vocabulary detector (*i.e.* ResNet-101 + instance level grounding head) on the synthetic images with base categories, and generate pseudo-labels for novel categories. When training detectors on such synthetic dataset, it gives 31.9 AP on novel categories on the COCO-OVD benchmark, 10.4 AP lower than tight coupling, showing the benefits of rich semantic and positional information encoded in SDM's visual features.

| $\mathcal{L}_{base}$ | $\mathcal{L}_{novel}$ | Detector AP | Precision | Recall |
|---|---|---|---|---|
| ✓ | | 70.2 | 87.9 | 68.3 |
| ✓ | ✓ | 79.7 | 89.1 | 90.0 |

Table S1. The quality of the pseudo-labels.

## S3. Quality of Pseudo-labels

Here we evaluate the quality of the pseudo-labels generated by the proposed grounding head. We adopt two metrics to assess their quality: (i) Detector AP and (ii) Precision and Recall. For Detector AP, we leverage the pre-trained Mask-RCNN model on the COCO dataset to generate ground truths (GTs) for the synthetic images, and then compute the AP of the pseudo labels derived from the teacher model. In the case of Precision and Recall, we randomly select and annotate 200 synthetic images, then calculate the precision and recall of their pseudo-labels. As shown in Table S1, after self-training on novel categories, the quality of the pseudo-labels can be significantly improved in terms of Detector AP (70.2%→79.7%), Precision (87.9%→89.1%) and Recall (68.3%→90.0%).

## S4. Qualitative Results

We show more qualitative results generated by our InstaGen in Figure S1. Without any manual annotations, InstaGen can generate high-quality images with object bounding-boxes of novel categories. In Figure S2, we further show the qualitative results predicted by the Faster R-CNN trained with the synthetic images form InstaGen on COCO validation set. The detector can now accurately localize and recognize the objects from novel categories.

| Method | Supervision | Detector | Backbone | Input Size | AP | $AP_c$ | $AP_f$ | $AP_{novel}$ |
|---|---|---|---|---|---|---|---|---|
| ViLD-ens. [3] | CLIP | Mask R-CNN | R50 | 1024×1024 | 25.5 | 24.6 | 30.3 | 16.6 |
| Detic [7] | CLIP | Mask R-CNN | R50 | 1024×1024 | 26.8 | 26.3 | 31.6 | 17.8 |
| F-VLM [5] | CLIP | Mask R-CNN | R50 | 1024×1024 | 24.2 | - | - | 18.6 |
| PromptDet [2] | CLIP | Mask R-CNN | R50 | 800×800 | 21.4 | 18.5 | 25.8 | 19.0 |
| DetPro [1] | CLIP | Mask R-CNN | R50 | 800×800 | 25.9 | 25.6 | 28.9 | 19.8 |
| BARON [6] | CLIP | Mask R-CNN | R50 | 800×800 | 25.1 | 24.4 | 28.9 | 18.0 |
| BARON [6]† | CLIP | Mask R-CNN | R50 | 800×800 | 27.6 | 27.6 | 29.8 | 22.6 |
| InstaGen | Stable Diffusion | Mask R-CNN | R50 | 800×800 | 23.0 | 20.6 | 27.1 | 20.3 |

Table S2. Results on open-vocabulary LVIS benchmark. † indicates using ensembling strategy for classification scores and learned prompts for the category's names.

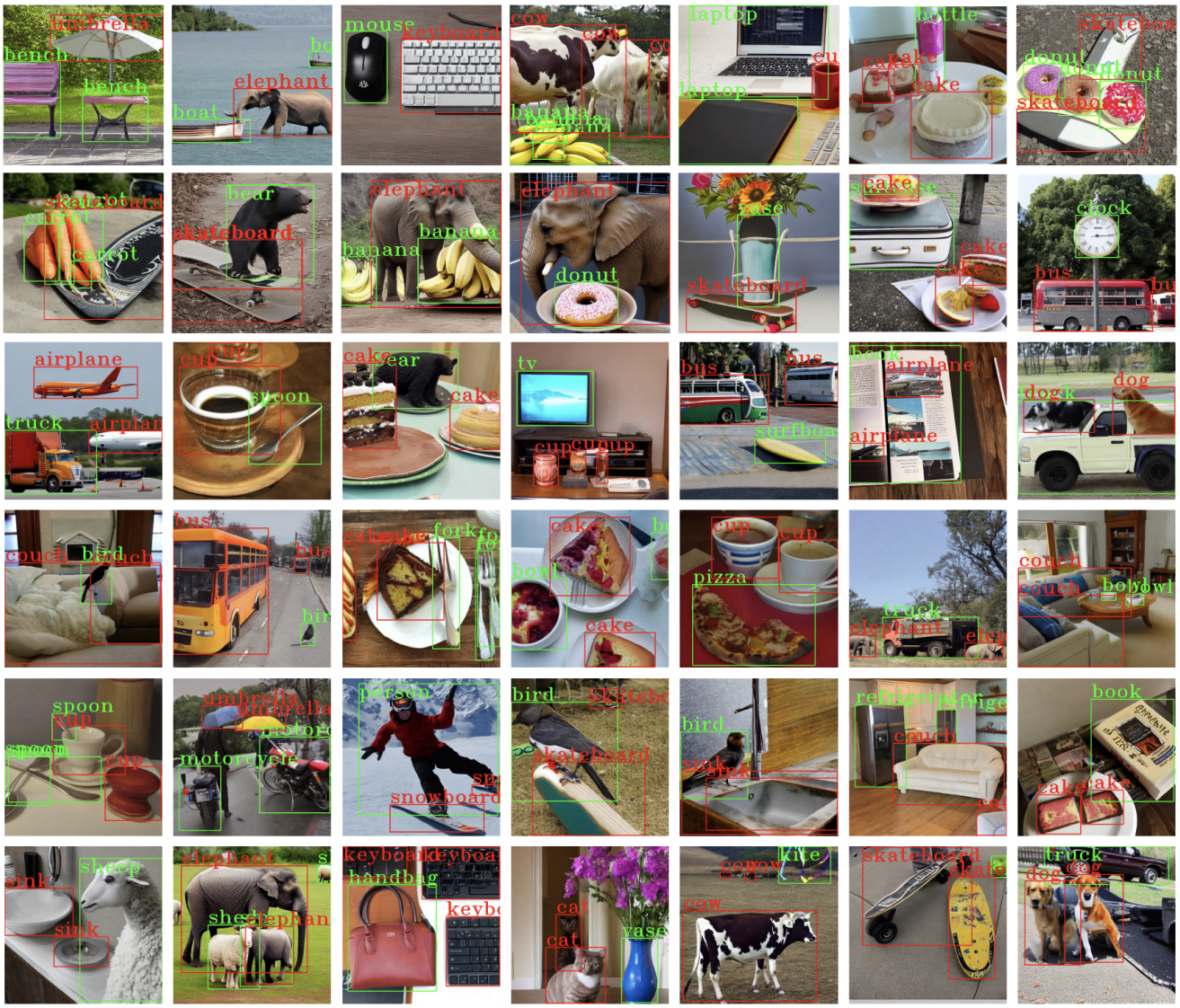

Figure S1. Qualitative results generated by our InstaGen. The bounding-boxes with green denote the objects from base categories, while the ones with red denote the objects from novel categories.

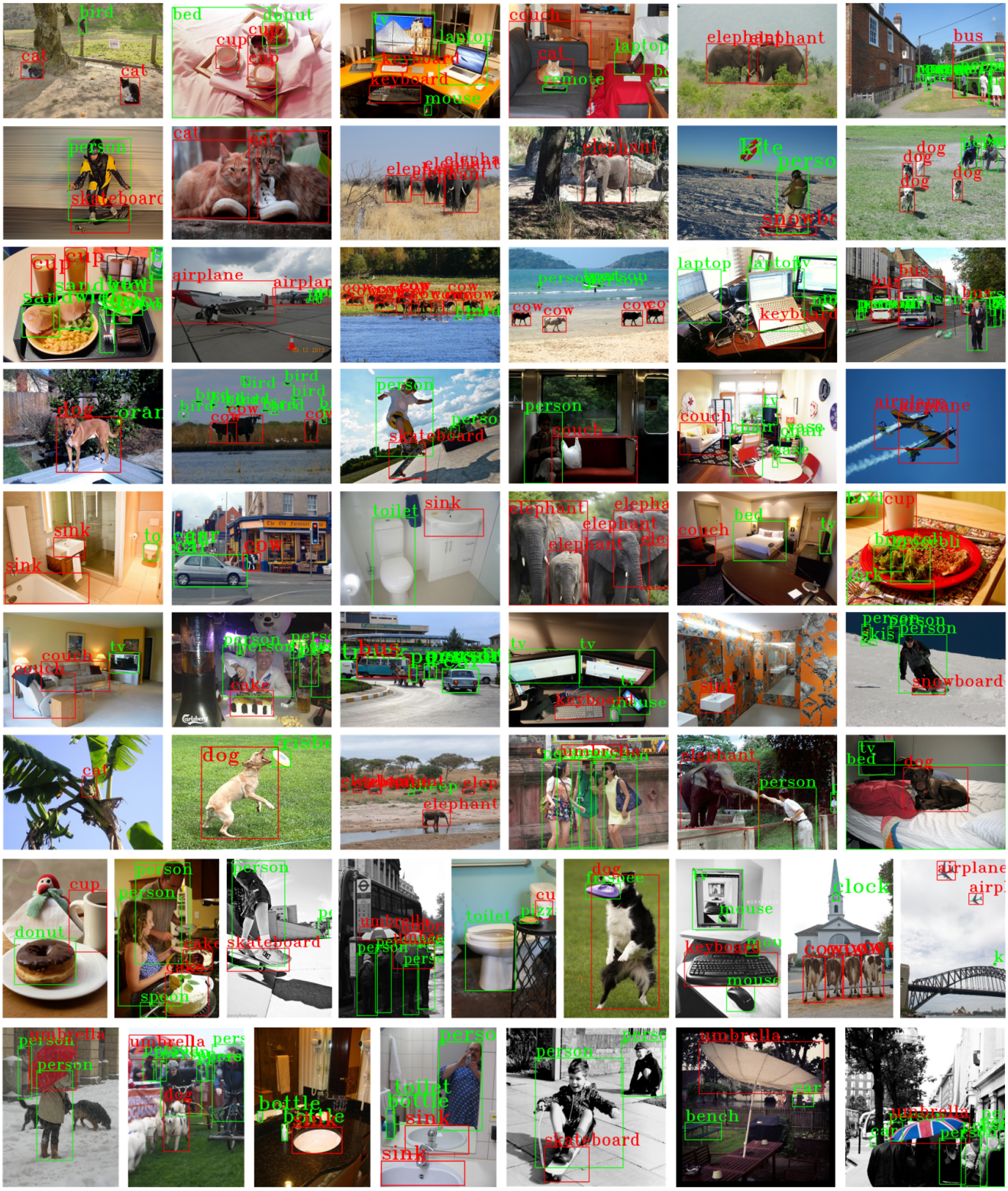

Figure S2. Qualitative results from our Faster R-CNN trained with the synthetic images from InstaGen on COCO validation set. The bounding-boxes with green denote the objects from base categories, while the ones with red denote the objects from novel categories.

## References


[1] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pages 14084–14093, 2022. 2

[2] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, pages 701–717. Springer, 2022. 1, 2

[3] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1, 2

[4] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 1

[5] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. 2022. 2

[6] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, pages 15254–15264, 2023. 2

[7] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. 1, 2