A. Extra Implementation Details

We set the exponential moving average factor as 0.99996. We follow prior works in the DETR literature to apply auxiliary loss on the output of each intermediate decoder layer, but anchor pre-matching is conducted only once. The hyper-parameters of the matching cost are identical to the corresponding loss coefficients. During training, we use random flip, random resize and random crop to augment the input image. The smaller edge of an input image is resized to a value between 480 and 800, and the larger edge is resized while keeping the aspect ratio. The maximum length of the resized larger edge is 1333. During inference, the temperature τ of the classification logits is set to 0.01. For image resolution during inference, the smaller edge is resized to 800, and the larger edge is resized accordingly by keeping the aspect ratio. The image is further resized when necessary to make sure the larger edge is no longer than 1333. During inference, we multiply the logit of novel classes by a factor of 8.0.

B. Localization Capability of CORA

In the main text, we demonstrate the effectiveness of our method by evaluating it on both region classification and object detection. In this section, we further show the superior novel object localization capability of our method.

In order to evaluate the localization capability, we only take the predicted box coordinates from the model for evaluation. The class label of each box is assigned by the ground truth box with highest IoU. The confidence score is replaced by the highest IoU. These modifications eliminate the effect of the classifier, and make sure that only the localization capability is evaluated. The predictions are then evaluated on the standard COCO OVD benchmark.

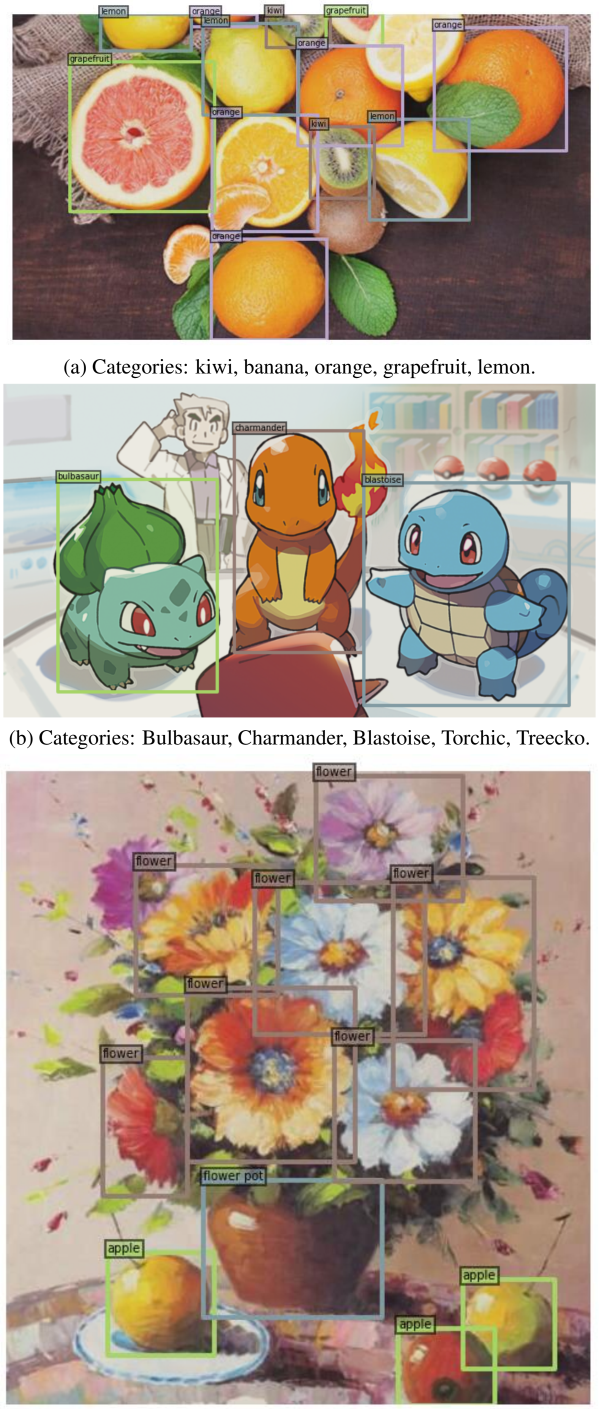
Method	Novel	Base
RegionCLIP	81.7	88.2
CORA	83.4	88.1

Table 6. Comparison on the localization capability. The performance is evaluated in AP50.

The result is shown in Tab. 6. We compare our method with RegionCLIP, which is a strong baseline on the COCO OVD benchmark. The novel-to-base performance gap of CORA is significantly lower than the baseline, demonstrating a better generalization capability towards the novel classes.

C. Visualizations

Fig. 5 visualizes the predictions of CORA on images with novel objects.



(a) Categories: kiwi, banana, orange, grapefruit, lemon.

(b) Categories: Bulbasaur, Charmander, Blastoise, Torchic, Treecko.

(c) Categories: flower, apple, flower pot.

Figure 5. Visualization of predictions on base and novel classes.