Appendix: Additional Visualization Results

Fig. 6: Additional visualization results on PASCAL VOC. "GT" denotes ground truth.

Here we present more qualitative results on PASCAL VOC [18] (Figure 6) and COCO-Object [5] (Figure 7), and compare our model with vanilla CLIP [43] and MaskCLIP [69]. As is shown, while our SCLIP model yields very clear segmentation masks in most cases, the vanilla CLIP model fails to correctly localize the primary objects within the images, and MaskCLIP often predicts with noticeable noise and many incoherent segments.

Specifically, in datasets with relatively fewer categories such as PASCAL VOC (see Figure 6), SCLIP is able to detect very detailed semantic features. For instance, in the first example, our model accurately segments the legs of the sheep although they occupy only a very small area in the image; and in the fourth example, our segmentation mask clearly displays the shape of the branches in the potted plants, which is slightly coarser than the ground truth but significantly outperforms the result of MaskCLIP, which categorizes the area around the potted plants along with the background as a single class. These observations testify the remarkable effectiveness of our CSA module.

Semantic segmentation with more categories (*e.g.*, 81 for COCO-Object) might be very challenging for zero-shot models. As is shown in Figure 7, in the absence of considering semantic correlations between the patch-level visual tokens, many noisy predictions emerge in MaskCLIP's segmentation results (*e.g.*, the first and the third examples). Notably, this issue cannot be simply addressed by leveraging additional refinement or thresholding strategies, since it may lead the model to segment the image into one or very few categories and thereby degrades its inference capabilities of detailed visual features. In addition, there are

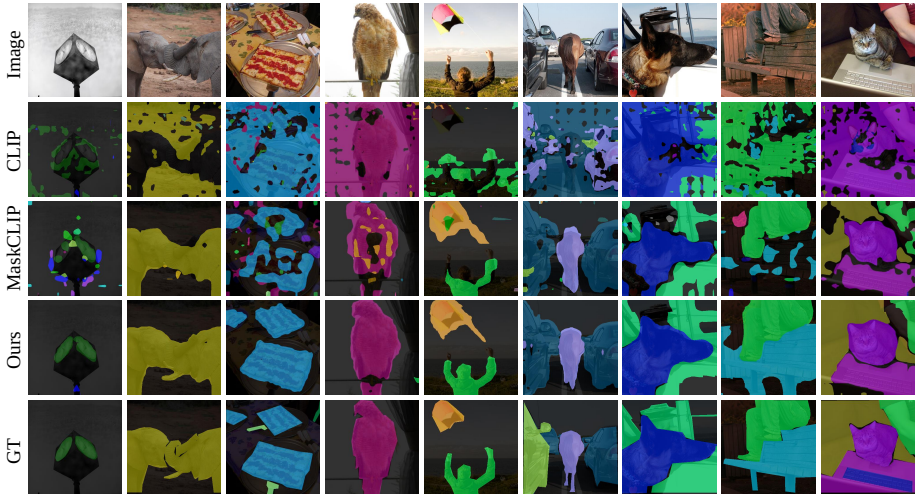

Fig. 7: Additional visualization results on COCO-Object. "GT" denotes ground truth.

several interesting observations such as in the fourth example, our segmentation mask of the bird skips the fence it stands on while the ground truth does not; and in the sixth example, the SUV on the left is annotated as "bus" while our model categories it to "car".