# Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models

Supplementary Material

## 7. Additional Qualitative Results

We provide additional qualitative results of images from Pascal Context and COCO Stuff in Fig 6. We note that, in the third image of the third row, PnP-OVSS correctly recognizes four loaves of bread whereas the ground truth only annotates one.

The rightmost column of the bottom three rows, separated from the rest by a red dash line, are examples of failure cases. The example on top contains multiple small instances of the same class (`people` and `surfboard`), which are understandably hard. The middle example contains multiple instances of `people`, which causes difficulties for PnP-OVSS to cover all objects. In addition, images with a clutter of different objects and complex texture (as in the last image) often cause a drop in performance.

## 8. Qualitative Results in the Wild

Fig 5 show qualitative results of PnP-OVSS+BLIP$_{\text{Flickr}}$, containing objects not seen in common semantic segmentation datasets, including cartoon characters `Minions` and `Gru`, dog breeds like `Samoyed` and `Border Collie`, food items like `Hamburger`, `Fries`, `Coke`, and `Fried Chicken`, places of interest like `Eiffel Tower` and `Merry-go-round`, a new electronic vehicle `Cybertruck`, and a celebrity `Elon Mask`.

In particular, we would like to point out the difficulty involved in segmenting the Steamboat Willie image, which is in greyscale and has little texture, depriving the network the ability to use color and texture features. Despite that, PnP-OVSS is able to extract the main components of `Mickey Mouse`, the `Helm` and the `Deck`.

## 9. PnP-OVSS Implementation Detail

**BLIP.** We use the ITM branch of BLIP_large, which adopts VIT-L/16 as the image encoder, BERT as the text encoder, and insert an extra cross-attention layer for each transformer block of BERT. For each cross-attention layer, the hidden size is 768, and the number of heads is 12. We interpolate the positional embedding to allow input resolution of 768 x 768. We adopt pretrained weights from two checkpoints for image retrieval on COCO and Flickr.

**BridgeTower.** We use the ITM branch of BridgeTower_*large*, which adopts VIT-L/14 as the image encoder, RoBERTa_*large* as the text encoder and an 6-layer cross attention encoder. For each cross attention layer of the cross-modal encoder, the hidden size is set to 1,024, and the number of heads is set to 16. We interpolate the positional embedding to allow input resolution of 770 x 770. The model weights are from the `bridgetowerlarge-itm-mlm-itc` checkpoint from Huggingface.

**Random Search.** We adopt the random search routine from the Gradient-Free-Optimizers library[1] [60] with our reward metric (§3.4). To parallelize the search process, we divide the search space into three groups and place each group on a GPU card. We perform 34 search iterations in each group. The best hyperparameter set from the three groups with the highest reward is taken as the final search result.

**Class Split for Densely Supervised Models.** We follow the most common setting [4, 18, 32, 47, 77, 90] which save pottedplant, sheep, sofa, train, tvmonitor as the 5 unseen classes for Pascal VOC; cow, motorbike, sofa, cat, boat, fence, bird, tvmonitor, keyboard, aeroplane as 10 unseen classes for Pascal Context; frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wallconcrete, tree, grass, river, clouds, playingfield, as 15 unseen classes for COCO Stuff.

## 10. Inference Speed

MaskClip, Reco, and PnP-OVSS take 0.05s, 4.41s, and 2.46s on average, respectively, for inference on a $320 \times 320$ image. We calculate the inference speeds of the three models on a single A6000 GPU with 48GB of RAM. The results are the average of 20 independent runs. Hence, PnP-OVSS achieves substantially better performance without significant increase in inference time.

## 11. Vil-Seg Evaluation Detail

In Tab 3, different from other methods that require weakly supervised finetuning on image-text data, Vil-Seg is evaluated on subset of datasets. Specifically, the author evaluate their method on 5 classes (potted plant, sheep, sofa, train, tv-monitor) out of the 20 object categories in PASCAL VOC; 4 classes (cow, motorbike, sofa, cat) out of the 59 object categories in PASCAL Context; and 15 classes (frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wall concrete, tree, grass, river, clouds,

[1] https://github.com/SimonBlanke/Gradient-Free-Optimizers

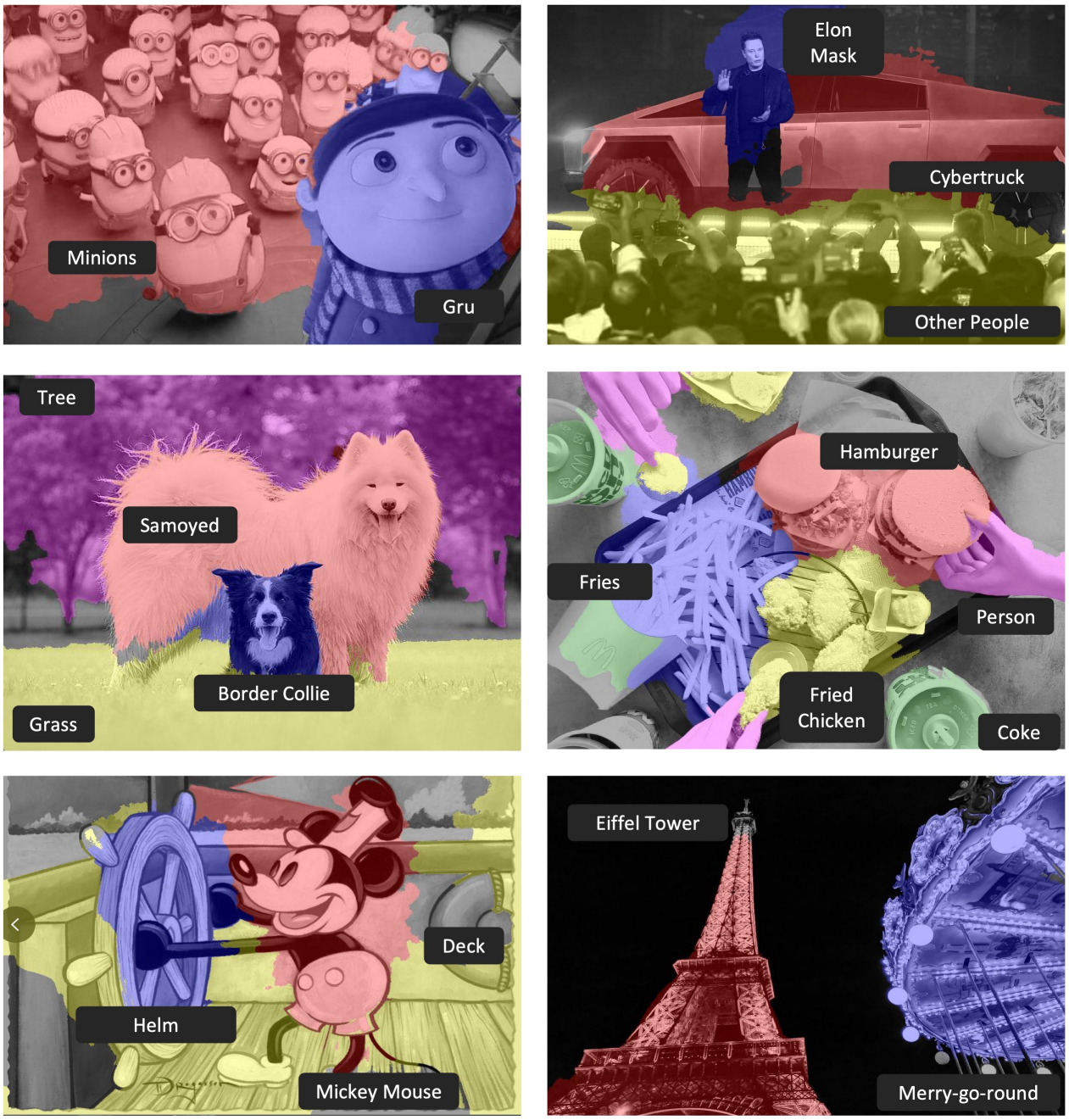


Figure 5. PnP-OVSS+BLIP$_{\text{Flickr}}$ segmentation result for in the wild images.

playing-field) out of the 171 object categories in COCO Stuff dataset.

## 12. Details of Zero-shot Semantic Segmentation Techniques

We summarize current methods for zero-shot semantic segmentation in Tab 7 to enable straightforward comparison between methods. Specifically, we include the supervision used, whether the method require pretraining and finetuning, the pretraining weight, finetuning data and total data size used in each method.

## 13. PnP-OVSS with ALBEF and mPLUG

As shown in Tab 8, we further apply PnP-OVSS on two other vision language models with cross-attention and image text matching loss, ALBEF [35] and mPLUG [34].


Figure 6. **Qualitative Results of PnP-OVSS + BLIP.** Images are from Pascal Context and COCO stuff. The bottom rows show the ground-truth (GT); the rest are our results. The last column of the last three rows, after the red dash line, shows failure cases.

Nevertheless, the performances are not as good as BLIP or BridgeTower. ALBEF [35] is pretrained with only 14M data with ViT-B whereas BLIP and BridgeTower are pretrained with 129M/404M data with ViT-L. We speculate that vision language models require sufficient numbers of parameters and pretraining data to acquire localization capability. mPLUG [34] is another vision language model pretrained with 14M data and ViT-L. However, mPLUG is trained for both image task and video task. With a relatively smaller amount of pretraining data than BLIP or BridgeTower, as well as image and video dual-modality objectives, mPLUG also does not perform well in image object localization.

| Method | Supervision | Training | PT weight | Finetuning data | PT/T data size |
|---|---|---|---|---|---|
| ***Methods that require finetuning on dense annotations w/o VL Models*** | | | | | |
| SPNet [72] | Pixel+Self | PT+FT | ImageNet | Pascal Voc/COCO stuff | 119K |
| ZS3Net [4] | Pixel+Self | PT+FT | - | Pascal Voc/ Pascal Context | 5K |
| CaGNet [18] | Pixel+Self | PT+FT | - | Pascal Voc/Pascal Context /COCO stuff | 123K |
| STRICT [47] | Pixel+Self | PT+FT | ImageNet | Pascal Voc/COCO stuff | 119K |
| ***Methods that require finetuning on dense annotations w/ VL Models*** | | | | | |
| SimBase [77] | Pixel+Self | PT+FT | Maskformer/FCN + CLIP | COCO stuff | 400.1M |
| LSeg [32] | Pixel+Self | PT+FT | ImageNet+CLIP | Pascal VOC/COCO/FSS | 401.3M |
| MaskCLIP+ [90] | Pixel+Self | PT+FT | CLIP | COCO stuff | 400.1M |
| ***Methods that require finetuning on image-text pairs*** | | | | | |
| OVSegmentor [75] | Text+Self | PT+FT | DINO+BERT | CC4M | 4.3M |
| Vil-Seg [43] | Text+Self | PT+FT | CLIP | CC12M | 412M |
| GroupVit* [44] | Text | T | - | CC3M+COCO | 3.4M |
| GroupVit [74] | Text | T | - | CC12M+YFCC14M | 26M |
| CLIPpy [51] | Text+Text | PT+FT | DINO+T5 Sentence | HQITP-134M | 134M |
| SegCLIP [44] | Text+Self | PT+FT | CLIP | CC3M+COCO | 403.4M |
| Viewco [52] | Text | PT+FT | GroupViT | CC12M+YFCC14M | 26M |
| TCL[6]+PAMR[2] | Text | PT+FT | CLIP | CC12M+CC3M | 415M |
| PACL [46] | Text | PT+FT | CLIP | GCC3M+GCC12M +YFCC15M | 430M |
| ***Methods that require finetuning but not image-text pair*** | | | | | |
| MaskCLIP [90] w/ ST | Text+ST | PT+T | CLIP | ImageNet1K | 401.2M |
| ZeroSeg* [8] | Text+Self | PT+T | CLIP | CC3M+COCO | 403.4M |
| ZeroSeg [8] | Text+Self | PT+T | CLIP | ImageNet1K | 401.2M |
| ***Methods that require no finetuning*** | | | | | |
| MaskCLIP [90] | Text | PT | CLIP | - | 400M |
| Reco[58] | Text | PT | CLIP+ImageNet | - | 400M |
| PnP-OVSS (Ours) | | | | | |
| + BLIP | Text | PT | BLIP_Flickr/BLIP_COCO | - | 129M |
| + BridgeTower | Text | PT | BridgeTower | - | 404M |

Table 7. Current methods for zero-shot semantic segmentation. Pixel represents method require pixel level annotation, Self represents method leverage self supervision, and Text represents method leverage image-text pair annotation. PT stands for Pre-training, T stands for training, ST stands for Self-training, FT stands for finetuning. All methods with pixel supervision are trained on seen categories and tested on unseen categories. For data size, We calculate only the image-caption data used for pretraining and all type of data for finetuning.

| Method | Training | HT on Dense Labels | Short-side Resolution | Pascal VOC-20 | Pascal Context-59 | COCO Object-80 | COCO Stuff-171 |
|---|---|---|---|---|---|---|---|
| PnP-OVSS (Ours) | | | | | | | |
| + ALBEF | × | × | 336 | 10.8 | 6.3 | 8.9 | 10.3 |
| + ALBEF | × | × | 768 | 11.1 | 6.8 | 8.8 | 10.7 |
| + mPLUG | × | × | 336 | 9.2 | 8.8 | 7.9 | 6.7 |

Table 8. Zero-shot semantic segmentation performance in mIoU.