# Supplementary Material for “Image Segmentation Using Text and Image Prompts”

## Experimental Setup

Throughout our experiments we use PyTorch [1] with CLIP ViT-B/16 [2]. We train on PhraseCut [3] for 20,000 iterations on batches of size 64 with an initial learning rate of 0.001 (for VitSeg 0.0001) which decays following a cosine learning rate schedule to 0.0001 (without warmup). We use automatic mixed precision and binary cross entropy as the only loss function.

## Image-size Dependency of CLIP

Since multi-head attention does not require a fixed number of tokens, the visual transformer of CLIP can handle inputs of arbitrary size. However, the publicly available CLIP models (ViT-B/16 and ViT-B/32) were trained on 224 × 224 pixel images. In this experiment we investigate how CLIP performance relates to the input image size – measured in a classification task. To this end, we extract the CLS token vector in the last layer from both CLIP models. Using this feature vector as an input, we train a logistic regression classifier on a subset of ImageNet [4] classes differentiating 67 classes of vehicles (Fig. 1). Our results indicate that CLIP generally handles large image sizes well, with the 16-px-patch version (ViT-B/16) showing a slightly better performance at an optimal image size of around 350 × 350 pixels.

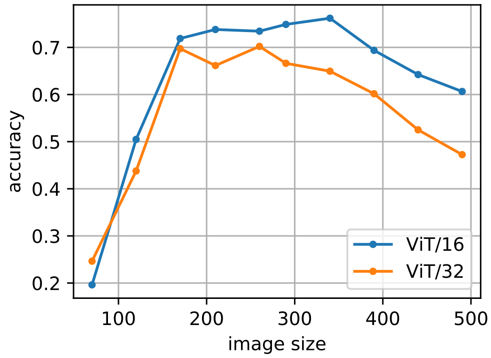

Figure 1. Image classification performance of CLIP over different image sizes.

## Object-mapping for Affordances and Attributes

For our systematic analysis on generalization (Section 5.5 in the main paper), we generate samples by replacing the following object categories by affordances (bold).

Affordances:
**sit on**: armchair, sofa, loveseat, deck chair, rocking chair, highchair, deck chair, folding chair, chair, recliner, wheelchair
**drink from**: bottle, beer bottle, water bottle, wine bottle, thermos bottle
**ride on**: horse, pony, motorcycle

Attributes:
**can fly**: eagle, jet plane, airplane, fighter jet, bird, duck, gull, owl, seabird, pigeon, goose, parakeet
**can be driven**: minivan, bus (vehicle), cab (taxi), jeep, ambulance, car (automobile)
**can swim**: duck, duckling, water scooter, penguin, boat, kayak, canoe

Meronymy (part-of relations):
**has wheels**: dirt bike, car (automobile), wheelchair, motorcycle, bicycle, cab (taxi), minivan, bus (vehicle), cab (taxi), jeep, ambulance
**has legs**: armchair, sofa, loveseat, deck chair, rocking chair, highchair, deck chair, folding chair, chair, recliner, wheelchair, horse, pony, eagle, bird, duck, gull, owl, seabird, pigeon, goose, parakeet, dog, cat, flamingo, penguin, cow, puppy, sheep, black sheep, ostrich, ram (animal), chicken (animal), person

## Average Precision Computation

The average precision metric has the advantage of not depending on a fixed threshold. This is particularly useful when new classes occur which lead to uncalibrated predictions. Instead of operating on bounding boxes as in detection, we compute average precision at the pixel-level. This makes the computation challenging, since AP is normally computed by sorting all predictions (hence all pixels) according their likelihood, which requires keeping them in the working memory. For pixels, this is not possible. To circumvent this, we define a fixed set of thresholds and aggregate statistics (true-positives, etc.) in each image. Finally, we sum up the statistics per threshold level and compute the precision-recall curve. Average precision, which is the area under the precision-recall curve is computed using Simpson integration.

## Qualitative Predictions

In Fig. 2 we show predictions of ViTSeg (PC), analogous to Fig. 4 of the main paper. In fact, ViTSeg trained with visual samples (PC+) shows worse performance. The predictions clearly indicate the deficits of an ImageNet-trained ViT backbone compared to CLIP: Details in the prompt are not reflected by the segmentation and a large number of false positives occur.

## Text prompts, object sizes and classes

To develop a better understanding of when our model performs well, we compare different text prompts (Fig. 3), object sizes (Fig. 4, left) and object classes (Fig. 4, right). This evaluation is conducted on a pre-trained CLIPSeg (PC+). In all cases we randomly sample different prompt forms during training. Here we assess the performance on 5,000 samples of the PhraseCut test set.

We see a small effect on performance for alternative prompt forms. In terms of object size there is a clear trend towards better performance on larger objects. Performance over different classes is fairly balanced.

## References

[1] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshops*, 2017.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[3] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

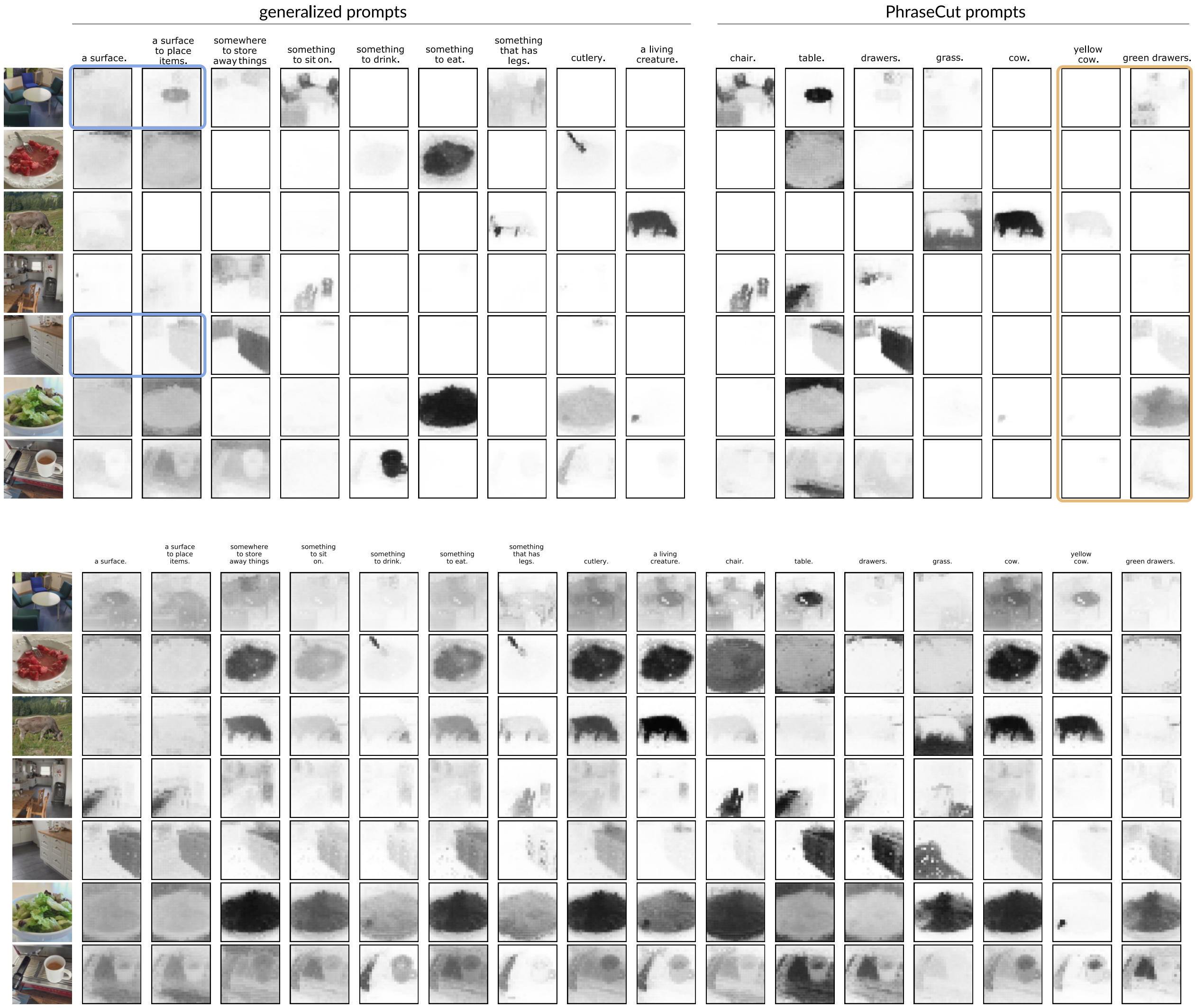

Figure 2. Qualitative predictions of CLIPSeg (PC+) (top, same as Fig. 4 of main paper for reference) and ViTSeg (PC) (bottom).

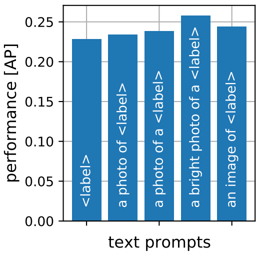

Figure 3. Effect of different text prompts on performance.

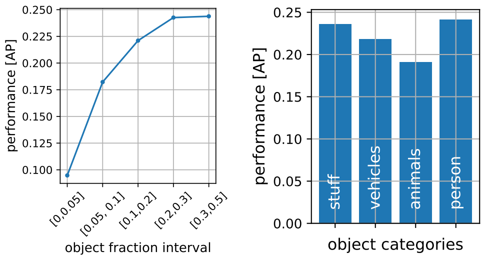

Figure 4. Effect of object size and class on performance.