Appendix: Image Segmentation Using Text and Image Prompts

Experimental Setup

Throughout our experiments we used PyTorch [1] and CLIP ViT-B/16 [2] weights. We trained 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 our only loss function.

Image-size Dependency of CLIP

Since multi-head attention does not require a fixed number of tokens, Transformers can handle inputs of arbitrary size. Nonetheless, does performance decrease when images of a different size than training size are used? To find out, we use CLIP as a feature extractor and using the CLS token in the last layer for training a logistic regression classifier. We do this on a subset of ImageNet [4] classes differentiating 67 classes of vehicles (Fig. 1).

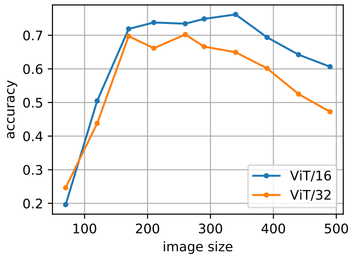

Figure 1: Image classification performance of CLIP over different image sizes.

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

Instead of operating on bounding boxes as in detection, we compute the metric 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 is computed using Simpson integration.

Qualitative Predictions

In Fig. 2 we show predictions of ViTSeg, analogous to Fig. 4 of the main paper. The predictions indicate the deficits of an ImageNet-trained ViT backbone compared to CLIP.

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.

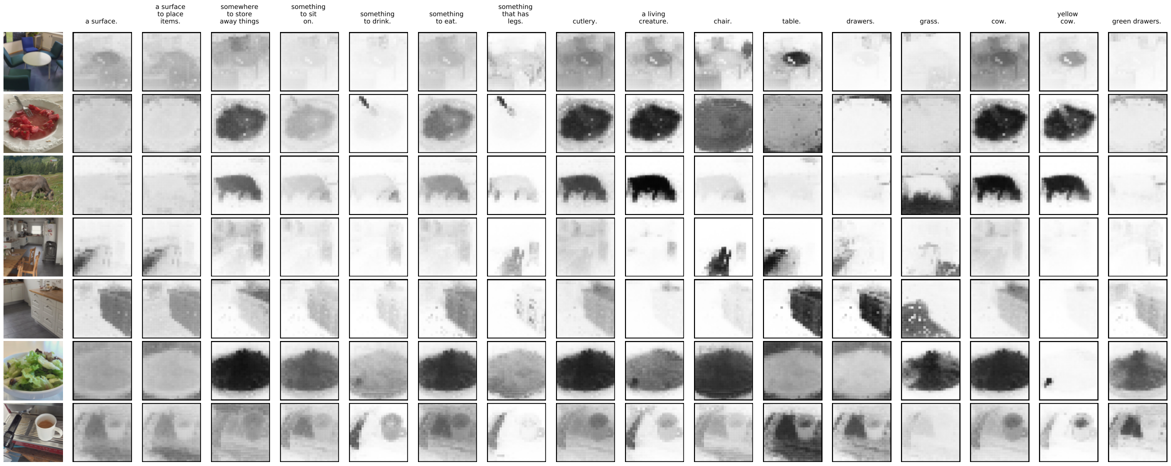

Figure 2: Qualitative predictions on the same data as Fig. 4.