

# Supplementary Material: Open-vocabulary Semantic Segmentation with Frozen Vision-Language Models

Chaofan Ma\*<sup>1</sup>

chaofanma@sjtu.edu.cn

Yuhuan Yang\*<sup>1</sup>

yangyuhuan@sjtu.edu.cn

Yanfeng Wang<sup>†1,2</sup>

wangyanfeng@sjtu.edu.cn

Ya Zhang<sup>1,2</sup>

ya\_zhang@sjtu.edu.cn

Weidi Xie<sup>1,2</sup>

weidi@sjtu.edu.cn

<sup>1</sup> Coop. Medianet Innovation Center,  
Shanghai Jiao Tong University, China

<sup>2</sup> Shanghai AI Laboratory

---

## Contents

<b>1</b>	<b>Datasets Settings</b>	<b>2</b>
<b>2</b>	<b>Prompt Templates</b>	<b>3</b>
<b>3</b>	<b>Ablations of Cross-modality Fusion Module</b>	<b>4</b>
<b>4</b>	<b>More Qualitative Visualization</b>	<b>5</b>

In this supplementary material, we start by giving a detailed description about our experimental setup, including datasets split, and prompt templates. The results for ablation study on the architecture of cross-modality fusion module are then presented. Lastly, more qualitative visualization results are displayed.

## 1 Datasets Settings

In Table 1, we provide the detailed datasets split settings used in our experiments. Datasets PASCAL-5<sup>i</sup> [10, 11] and COCO-20<sup>i</sup> [12] follow the split settings proposed in LSeg [13]. COCO-20<sup>i</sup> → PASCAL VOC refers to the transferability experiment mentioned in Section 4.6 of the main paper.

Dataset	fold 0	fold 1	fold 2	fold 3
PASCAL-5 <sup>i</sup>	aeroplane, bicycle, bird, boat, bottle	bus, car, cat, chair, cow	diningtable, dog, horse, motorbike, person	pottedplant, sheep, sofa, train, tv/monitor
COCO-20 <sup>i</sup>	person, aeroplane, boat, parkingmeter, dog, elephant, backpack, suitcase, sportsball, skateboard, wineglass, spoon, sandwich, hotdog, chair, diningtable, mouse, microwave, refrigerator, scissors	bicycle, bus, trafficlight, bench, horse, bear, umbrella, frisbee, kite, surfboard, cup, bowl, orange, pizza, sofa, toilet, remote, oven, book, teddybear	car, train, firehydrant, bird, sheep, zebra, handbag, skis, baseballbat, tennisracket, fork, banana, broccoli, donut, pottedplant, tvmonitor, keyboard, toaster, clock, hairdrier	motorbike, truck, stopsign, cat, cow, giraffe, tie, snowboard, baseballglove, bottle, knife, apple, carrot, cake, bed, laptop, cellphone, sink, vase, toothbrush
COCO-20 <sup>i</sup> → PASCAL VOC	aeroplane, boat, chair, diningtable, dog, person	bicycle, bus, horse, sofa	bird, car, pottedplant, sheep, train, tvmonitor	bottle, cat, cow, motorbike

**Table 1.** Data split for PASCAL-5<sup>i</sup>, COCO-20<sup>i</sup>, and COCO-20<sup>i</sup> → PASCAL VOC. PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup> use 4-fold cross-validation, and categories in each row are for test and the rest classes are used for training. In COCO-20<sup>i</sup> → PASCAL VOC, categories in fold  $i$  are novel classes in PASCAL VOC after removing the seen classes in corresponding training split on fold  $i$  of COCO-20<sup>i</sup>.

## 2 Prompt Templates

Before feeding the semantic category into text encoder, we use multiple prompt templates as decorations to generate text embeddings for one category, and then ensemble these embeddings by averging. The following are the prompt templates we used:

```
'a bad photo of the {category}.',
'a photo of the large {category}.',
'a photo of the small {category}.',
'a cropped photo of a {category}.',
'This is a photo of a {category}',
'This is a photo of a small {category}',
'This is a photo of a medium {category}',
'This is a photo of a large {category}',
'This is a masked photo of a {category}',
'This is a masked photo of a small {category}',
'This is a masked photo of a medium {category}',
'This is a masked photo of a large {category}',
'This is a cropped photo of a {category}',
'This is a cropped photo of a small {category}',
'This is a cropped photo of a medium {category}',
'This is a cropped photo of a large {category}',
'A photo of a {category} in the scene',
'a bad photo of the {category} in the scene',
'a photo of the large {category} in the scene',
'a photo of the small {category} in the scene',
'a cropped photo of a {category} in the scene',
'a photo of a masked {category} in the scene',
'There is a {category} in the scene',
'There is the {category} in the scene',
'This is a {category} in the scene',
'This is the {category} in the scene',
'This is one {category} in the scene',
'There is a masked {category} in the scene',
'There is the masked {category} in the scene',
'This is a masked {category} in the scene',
'This is the masked {category} in the scene',
'This is one masked {category} in the scene',
```

### 3 Ablations of Cross-modality Fusion Module

We conduct an ablation study on the architecture of the cross-modality fusion module, *i.e.*, the number of layers, heads, and width of the transformer. The width/head ratio is set to 64 in this ablation. In Table 2, we can see that, a lightweight fusion module (1-layer transformer) already works reasonably well, and a steady improvement can be further observed as the number of the transformer layer increases from 1 to 12. However, performance begins to drop when the number of layers exceeds 12. Likewise, with a fixed number of layer, the performance variation on the widths (or heads) of the model follows the same scheme. Based on this result, we set layers to 12, and heads to 8 in all experiments in our main paper.

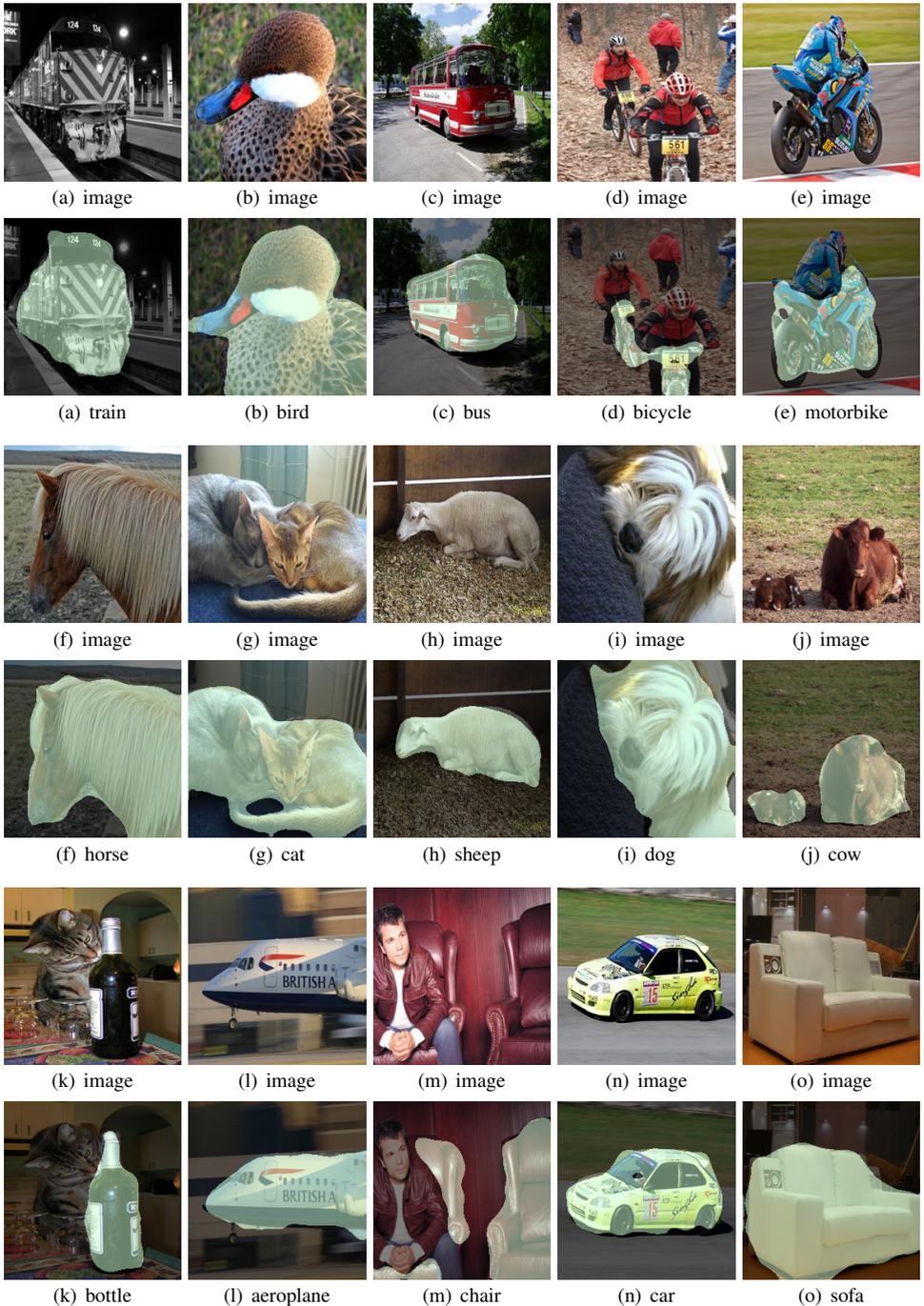
layers	heads	width	5 <sup>0</sup>	5 <sup>1</sup>	5 <sup>2</sup>	5 <sup>3</sup>	mIoU
1	8	512	52.2	59.2	45.0	44.6	50.2
3	8	512	57.1	68.0	46.7	50.0	55.4
6	8	512	58.9	68.9	<b>51.3</b>	51.5	57.7
<u>12</u>	<u>8</u>	<u>512</u>	<b>61.9</b>	70.0	51.2	<b>52.7</b>	<b>59.0</b>
18	8	512	57.0	69.5	43.5	49.4	54.8
24	8	512	51.1	<b>71.7</b>	48.7	47.2	54.7
12	1	64	54.7	56.3	41.4	43.8	49.1
12	4	256	60.9	65.9	43.2	45.9	54.0
<u>12</u>	<u>8</u>	<u>512</u>	<b>61.9</b>	70.0	<b>51.2</b>	<b>52.7</b>	<b>59.0</b>
12	10	640	54.1	<b>70.5</b>	49.1	47.7	55.3
12	12	768	55.5	65.8	47.8	46.8	54.0
12	16	1024	51.7	66.4	49.9	49.7	54.4

**Table 2.** Performance on layers, heads, and width of cross-modality fusion module on PASCAL-5<sup>i</sup>. Underlined are the best parameters we used in the main paper.

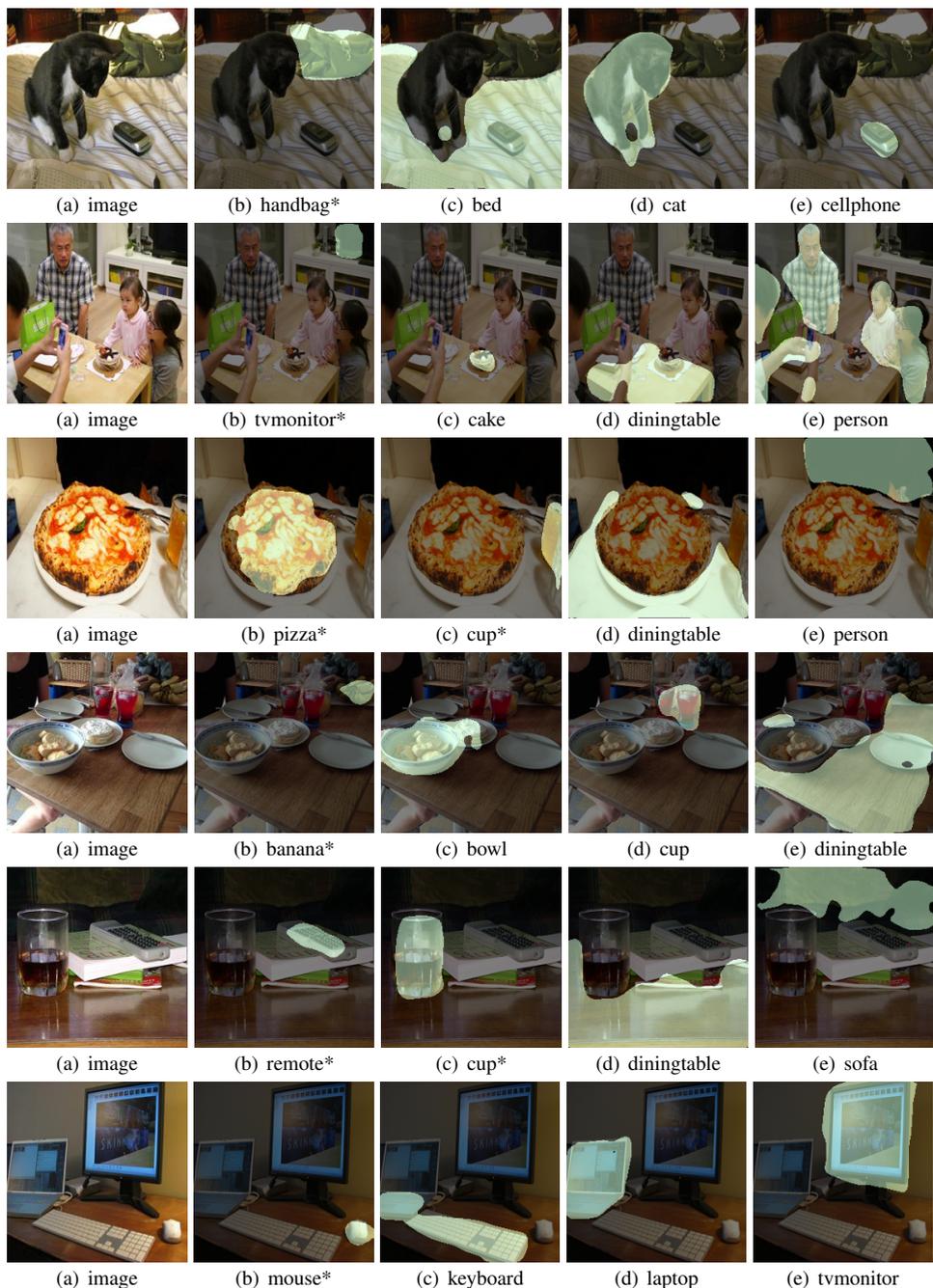
## 4 More Qualitative Visualization

In this section, we present additional qualitative results on PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup> using our model with ViT-L backbone. Specifically, Figure 1 shows the results on PASCAL-5<sup>i</sup>. All categories are novel (unseen) in their corresponding fold. Taking into account the variety of images, we choose 15 different categories to display. Figure 1(d) bicycle, Figure 1(e) motorbike, Figure 1(m) chair, and Figure 1(k) bottle show **Fusioner**'s ability to distinguish object of target semantic from other objects (distractor), *e.g.*, the person in Figure 1(e), and the cat in Figure 1(g). Besides, in Figure 1(d) bicycle, Figure 1(j) cow, Figure 1(m) chair, and Figure 1(g) cat, the model segments precisely even when the target instance contains more than one.

The visualization of COCO-20<sup>i</sup> is shown in Figure 2, with both seen and unseen categories are displayed. Facing a more noisy and complex scene, **Fusioner** is still able to recognize the desired (unseen) categories that are small and cornered, for example, tvmonitor, banana, remote and mouse in column of Figure 2(b).



**Figure 1.** Qualitative results on PASCAL5<sup>i</sup>. All categories are novel (unseen) in their corresponding folds, and we show 15 different categories. For each two rows, top: images; bottom: segmentation masks for unseen categories.



**Figure 2.** Qualitative results on COCO-20<sup>17</sup>. (a) input images, (b)-(e) segmentation masks for different categories. Unseen categories are marked as \*.

## References

- [1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [2] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 991–998. IEEE, 2011.
- [3] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.
- [4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.