Supplementary Material for
Pay Attention to Your Neighbours:
Training-Free Open-Vocabulary Semantic Segmentation

A. Gaussian kernel's standard deviation

In a realistic *training-free* open-vocabulary scenario, where additional data access is restricted, there should be no validation set available for hyperparameter tuning. Consequently, it is crucial for *training-free* methods to operate effectively without such procedures. In our approach, we introduce a hyperparameter denoted as σ, representing the standard deviation of the Gaussian kernel used in Eq. (10), which we set to 5 in our experiments. In this section, we detail the heuristics guiding this choice, enabling us to determine this value without the need for fine-tuning.

For a patch located at $\boldsymbol{\mu}$, the Gaussian kernel increases its attention logits by 1 at $\boldsymbol{\mu}$ and by lesser values at neighbouring patch locations. Our choice of σ is based on the number of neighbouring patches whose attention logits are modified by more than a threshold τ. To achieve this, we express:

$$\phi(\mathbf{x};\boldsymbol{\mu},\sigma) = \exp\left(-\frac{||\mathbf{x}-\boldsymbol{\mu}||^2}{2\sigma^2}\right) > \tau \tag{17}$$

$$\Rightarrow \quad \frac{||\mathbf{x}-\boldsymbol{\mu}||^2}{2\sigma^2} < -\ln\tau \tag{18}$$

$$\Rightarrow \quad ||\mathbf{x}-\boldsymbol{\mu}||^2 < -2\sigma^2\ln\tau \tag{19}$$

$$\Rightarrow \quad ||\mathbf{x}-\boldsymbol{\mu}||^2 < \left(\sigma\sqrt{-2\ln\tau}\right)^2. \tag{20}$$

Considering Eq. (20), neighbouring patches for which $\boldsymbol{\mu}$'s attention logits are increased by at least τ are positioned within of a circle centered on $\boldsymbol{\mu}$ with a radius of $\sigma\sqrt{-2\ln\tau}$. For instance, with $\sigma = 5$, patch $\boldsymbol{\mu}$'s attention to 37 patches gets a logit increase of at least 0.8 as illustrated in Fig. 4. Table 5 displays the value of this heuristic measure for $\sigma \in \{1, 2, \ldots, 10\}$ and $\tau \in \{0.7, 0.8, 0.9\}$. Besides, CLIP [32] has been trained on 224×224 pixel images, meaning the ViT-B/16 backbone operates on 14×14 patches for each image. Based on this fact and considering the values provided in Tab. 5, we opted for $\sigma = 5$ in our experiments to maintain a balance, *i.e.*, to have neither too small nor too large field of attention. It is worth noting that τ is defined solely for the purpose of the described heuristic and does not play a role in our approach. In other words, there is no τ value to fine-tune in our approach.

Although we employed a heuristic measure to determine σ, we provide in Fig. 5 the impact of varying σ values on test set performance. Please note that these experiments were conducted after deciding to use $\sigma = 5$, and whose

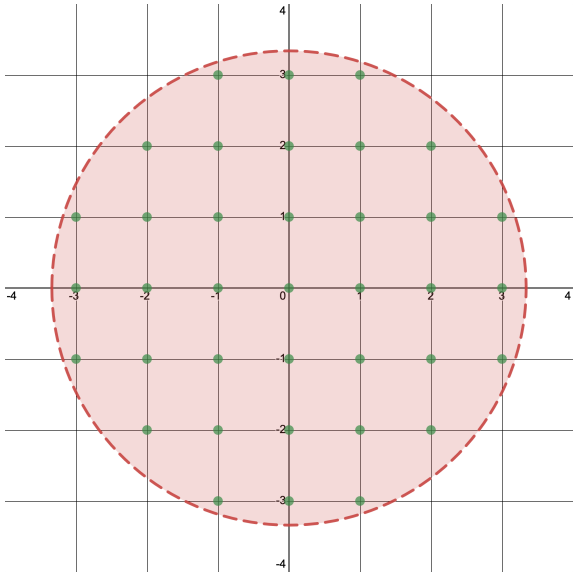

Figure 4. **Illustrative example of Eq. (20).** The attention logits of the center point to the points within the depicted circle are increased by at least τ. Example generated for $\sigma = 5$ and $\tau = 0.8$.

Table 5. **Proposed heuristic measure for determining σ value.** For 3 values of τ and 10 values of σ, the table provides the number of patches that patch $\boldsymbol{\mu}$'s attention logits to them is increased by more than τ. It is important to note that these values are derived based on an infinite grid of patches, while in practice, these numbers could be less depending on the window size and $\boldsymbol{\mu}$'s position.

σ	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
1	1	1	1
2	9	5	1
3	21	13	5
4	37	21	9
5	57	37	21
6	81	49	21
7	109	69	37
8	145	89	45
9	177	113	57
10	221	137	69

goal is to demonstrate that *i)* our heuristic approach to find σ provides indeed a good value; and *ii)* the performance across different datasets is not strongly sensitive to the hyperparameter σ.

B. Visual examples

Additional visual examples can be found in Fig. 6 for PASCAL Context (59) [29], and in Fig. 7 for COCO-Object [5, 23]. Upon reviewing the images in Fig. 6, we can observe that SCLIP [38] often encounters difficulties in segmenting objects wholly and finding their boundaries (*e.g.*, rows 1, 2, 4, and 8). We attribute this problem to SCLIP's failure to consistently incorporate information from surrounding patches. Similar observations can be made for the first four rows of Fig. 7. However, an interesting minute distinction between the methods emerges in the final row of the figure. Notably, the pixels representing the cat's eyes differ significantly from those of its skin, resulting in SCLIP failing to segment them as the same class. In contrast, NACLIP attentively considers the surrounding context of the eyes, resulting in accurate segmentation.

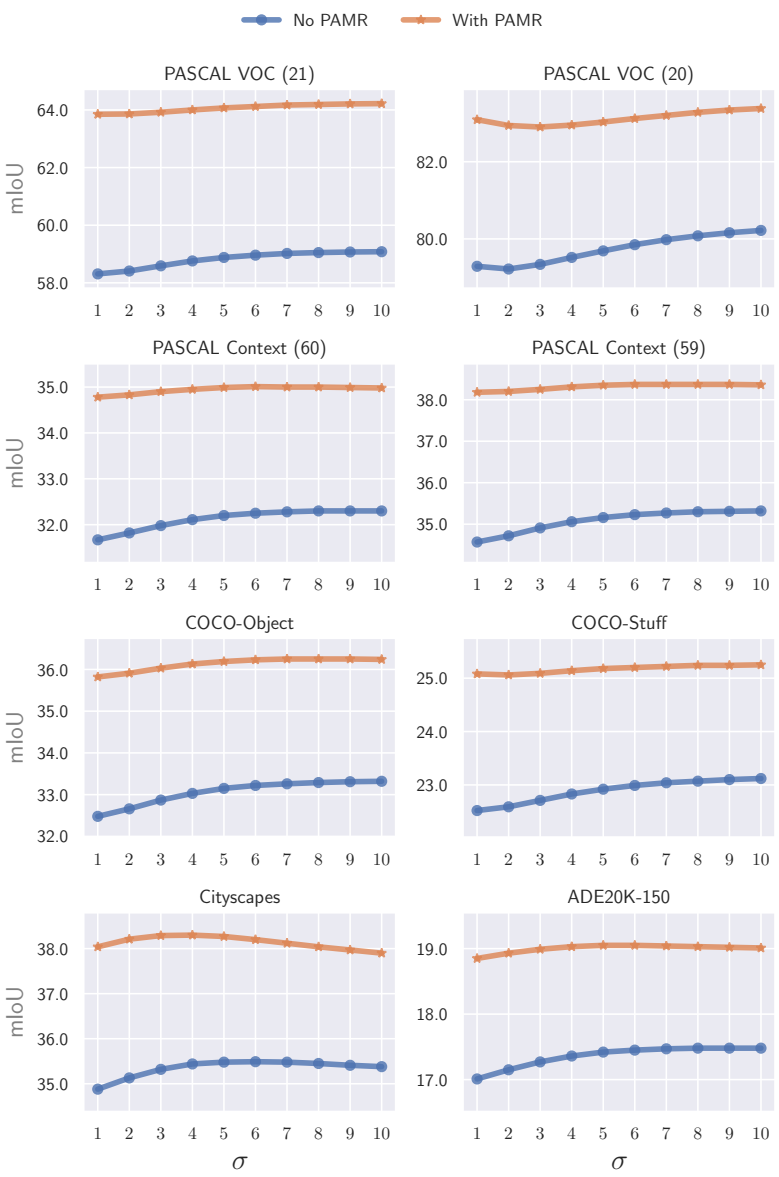

Figure 5. **Ablation study on the impact of σ.** We have provided results for both cases of using and not using post-processing, revealing consistent trends across both cases.

Figure 6. **Additional visual examples (segmentation maps) from PASCAL Context (59) [29]** for CLIP [32], SCLIP [38], and our method.

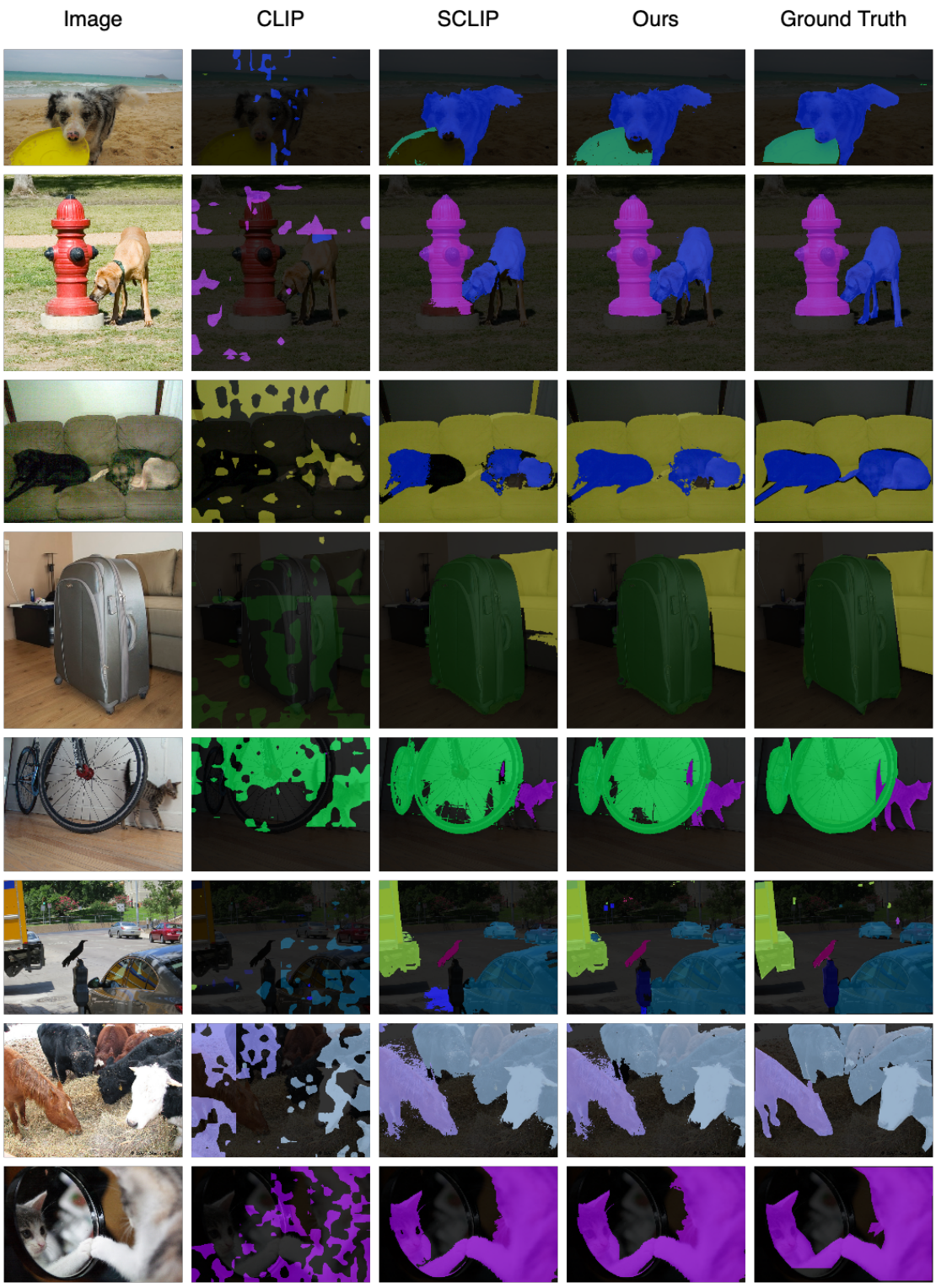

Figure 7. **Additional visual examples (segmentation maps) from COCO-Object [5, 23]** for CLIP [32], SCLIP [38], and our method.