A. Appendix

A.1. Additional Implementation Details

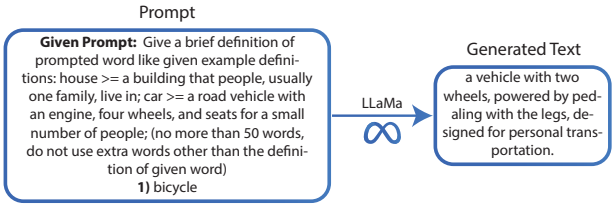
LLM-based Auxiliary Text Generation. As detailed in Sec. 3.4, we leverage LLama 3 [19] to generate auxiliary texts for each class name. We use generated definitions as the auxiliary text type for the COCO-Stuff [8], Pascal VOC [20], and Pascal Context [47] datasets and synonyms for the COCO-Object [40] and Cityscapes [12] datasets. Fig. 4a and Fig. 4b illustrate the procedure for generating definitions and synonyms, respectively. Example definitions and synonyms from the Cambridge Dictionary [9] are utilized to guide LLaMa in producing more precise definitions.

Image Engineering (λ) and Auxiliary Text Coefficients (α). The Image Engineering (λ) and Auxiliary Text Coefficients (α) are employed in weighted summations within the Image Engineering and LLM-based Auxiliary Text Generation modules, respectively. Tab. 9 reports these hyperparameter values used across all evaluated datasets. As demonstrated in Tab. 8, the effects of λ and α are insignificant within the range specified in the main paper, and our default values exhibit minimal variance across datasets.

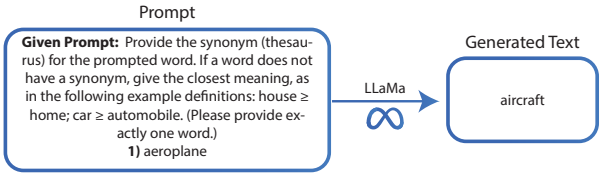
Background set. Following previous studies [42, 58, 61], we define a background set for the Pascal VOC and COCO-Object datasets to enable our method to distinguish foreground classes from the background. Specifically, the background set employed in COCO-Object is

`background =` [*sky, wall, tree, wood, grass, road, sea, river, mountain, sands, desk, bed, building, cloud, lamp, door, window, wardrobe, ceiling, shelf, curtain, stair, floor, hill, rail, fence*].

A similar background set is used for the evaluation of Pascal VOC.



(a) We employ the illustrated prompt to generate definitions.



(b) We employ the illustrated prompt to generate synonyms.

Figure 4. **Procedure for generating auxiliary texts for a given class name.**

Hyperparameter	Stuff	Object	VOC	Context	City
λ	0.75	0.75	0.7	0.75	0.7
α	0.2	0.1	0.05	0.15	0.05

Table 9. **Hyperparameter values used in our experiments.**

Backbone	VOC
ViT-L/14	53.3
ViT-B/32	56.7
ViT-B/16	**67.9**

Table 10. **Impact of different visual backbones.** We compare the performance of ITACLIP with different visual backbones.

Method	Background Set	Object
ITACLIP	✗	34.5
ITACLIP	✓	**37.7**

Table 11. **Effect of background set.** ITACLIP performs better when the background set is employed.

A.2. Additional Experiments

Impact of different visual backbones. We perform an ablation study to assess the impact of various CLIP-ViT backbones, as shown in Tab. 10. ITACLIP achieves peak performance using the ViT-B/16 backbone, consistent with prior works [25, 33].

Background set. In Tab. 11, we analyze the effect of defining the background set on the COCO-Object dataset. Since the COCO-Object dataset consists of 80 "thing" classes and one explicit background class, our method fails to distinguish foreground classes from the background when the word "background" is solely used to define all possible background classes. We observe a substantial performance boost when a separate background set is defined.

A.3. More Qualitative Results

Fig. 5 presents additional visualizations of ITACLIP on the COCO-Object, Pascal Context, and Pascal VOC datasets, comparing our method with SCLIP [61] and NACLIP [25]. As shown in Fig. 5, ITACLIP produces clearer segmentation masks compared to SCLIP and NACLIP, whose predictions are generally noisier. Furthermore, SCLIP and NACLIP occasionally fail to recognize objects accurately and predict classes not present in the image.

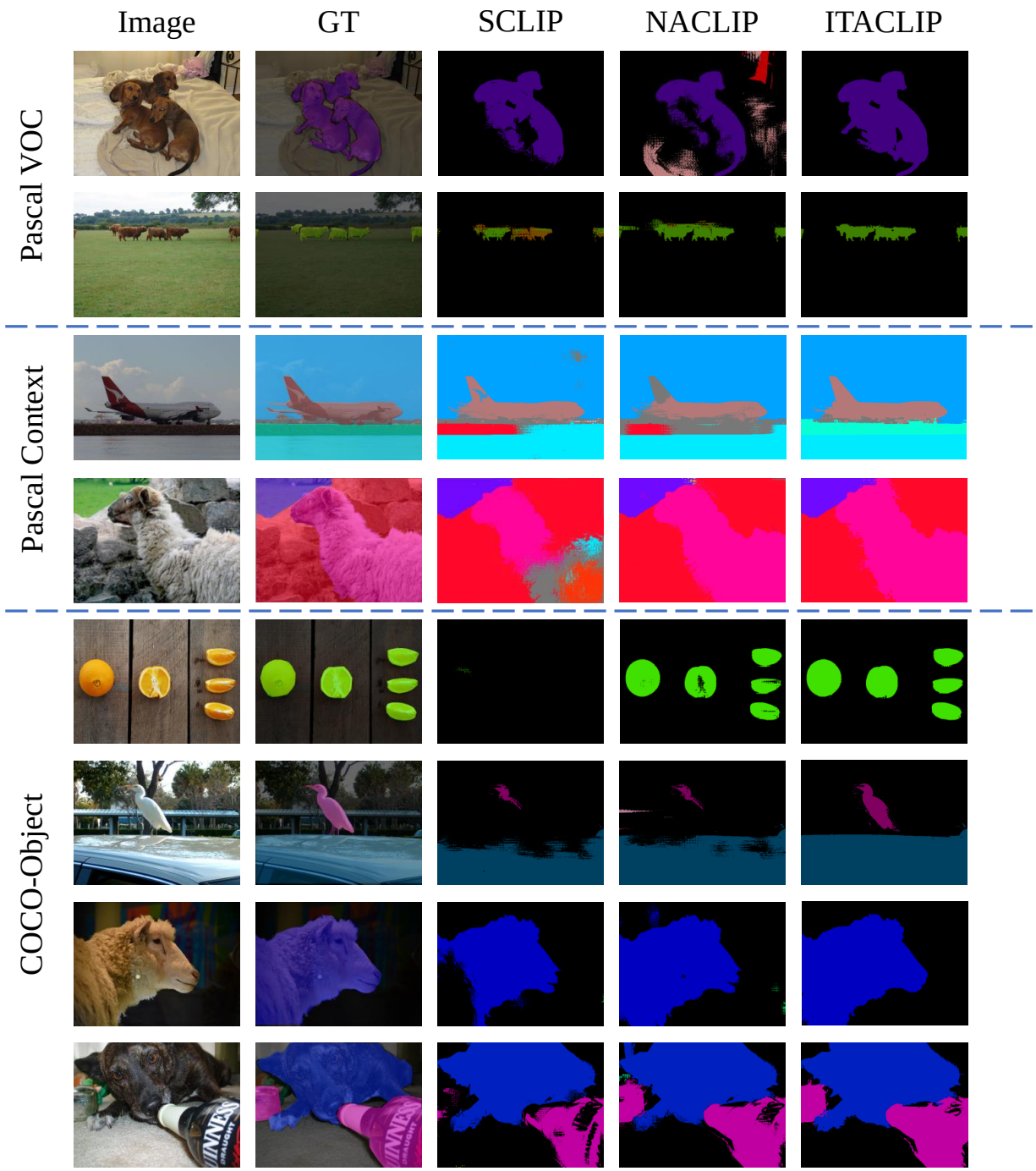


Figure 5. **Qualitative comparison of training-free semantic segmentation methods.** We compare ITACLIP with SCLIP [61] and NACLIP [25] using images from the Pascal VOC [20], Pascal Context [47], and COCO-Object [40] datasets. ITACLIP consistently outperforms the other approaches. GT denotes the ground truth of the image.