Table 1. Performance Comparison Across CLIP Architectures and Post-processing Methods. Comparative evaluation of our approach against state-of-the-art methods (Trident and CLIPer) across different CLIP architectures (CLIP-ViT-B/16, CLIP-ViT-L/14, OpenCLIP-ViT-H/14) with and without post-processing. Best results are shown in bold, second-best results are underlined. Note: CorrCLIP is not included in the comparison as its implementation is not publicly available.

Method	Post-processing	With background			Without background					Avg.
		VOC21	Context60	COCO-Obj	VOC20	City	Context59	ADE	Stuff	
Without Post-processing										
CLIP-ViT-B/16										
CLIPer(Sun et al., 2024a)	NO	60.1	34.8	36.0	84.0	-	38.5	19.8	25.3	-
Trident(Shi et al., 2024)	NO	64.5	37.2	**39.5**	83.7	40.4	40.9	20.9	**27.6**	44.5
Ours	NO	**66.1**	**37.7**	39.4	**85.4**	**42.8**	**41.2**	**21.0**	27.3	**44.8**
CLIP-ViT-L/14										
CLIPer(Sun et al., 2024a)	NO	61.2	34.3	39.6	**88.2**	-	39.8	21.8	25.8	-
Trident(Shi et al., 2024)	NO	61.4	36.4	40.2	84.8	40.4	39.8	**23.2**	26.4	42.5
Ours	NO	**67.0**	**37.3**	**40.2**	85.5	**42.7**	**41.0**	22.9	**27.1**	**45.5**
OpenCLIP-ViT-H/14										
CLIPer(Sun et al., 2024a)	NO	58.0	34.1	39.2	85.8	-	36.9	22.1	25.2	-
Trident(Shi et al., 2024)	NO	68.6	38.2	**40.8**	**87.7**	**43.6**	42.6	**25.4**	28.0	46.6
Ours	NO	**69.1**	**39.0**	40.0	86.9	43.0	**42.8**	24.5	**28.1**	**46.7**
With Post-processing										
CLIP-ViT-B/16										
CLIPer(Sun et al., 2024a)	YES	65.9	37.6	39.0	85.2	-	41.7	21.2	27.5	-
Trident(Shi et al., 2024)	YES	67.1	38.6	**41.1**	84.5	42.9	42.2	21.9	28.3	45.8
Ours	YES	**70.4**	**39.8**	40.8	**86.2**	**46.4**	**43.6**	**22.2**	**28.7**	**47.3**
CLIP-ViT-L/14										
CLIPer(Sun et al., 2024a)	YES	69.8	38.0	**43.3**	**90.0**	-	**43.6**	**24.4**	**28.7**	-
Trident(Shi et al., 2024)	YES	62.6	37.3	40.5	85.5	43.0	40.9	24.0	27.1	44.3
Ours	YES	**71.3**	**39.6**	40.7	86.5	**46.4**	43.3	23.8	28.5	**47.5**
OpenCLIP-ViT-H/14										
CLIPer(Sun et al., 2024a)	YES	**88.9**	39.3	**42.8**	**88.8**	-	43.2	24.4	28.3	-
Trident(Shi et al., 2024)	YES	70.8	40.1	42.2	88.7	**47.6**	44.3	**26.7**	28.6	**48.6**
Ours	YES	71.7	**40.5**	41.4	87.6	47.0	**44.8**	25.6	**28.8**	48.4

Table 2. Ablation Study on Different VFM Feature Extractors. We compare four DINO variants: ViT-Base with patch sizes of 8 (B8) and 16 (B16), and ViT-Small with patch sizes of 8 (S8) and 16 (S16); two DINOV2 variants: ViT-B/14 and ViT-S/14; and the ViT-B/16 architecture from SAM. Results demonstrate that DINOV2 ViT-S/14 achieves the best performance on the VOC20 dataset, while DINO ViT-B/8 outperforms all other architectures across the remaining datasets.

Model	Dataset							
	VOC21	Context-60	COCO	VOC20	Cityscapes	Context-59	ADE20K	COCO-Stuff
DINO								
ViT-B/8	**70.41**	**39.82**	**40.80**	86.18	**46.41**	**43.56**	**22.24**	**28.72**
ViT-B/16	69.26	39.67	40.46	85.99	45.56	42.47	22.11	27.88
ViT-S/8	69.96	39.49	40.63	86.02	45.94	42.79	22.17	28.24
ViT-S/16	68.91	39.12	40.27	85.88	45.23	42.01	21.96	27.42
DINOV2								
ViT-S/14	68.26	39.13	40.66	**86.61**	44.33	43.09	21.78	28.41
ViT-B/14	68.17	39.03	40.57	86.50	44.59	43.07	21.87	28.48
SAM								
ViT-B/16	68.36	38.0	38.84	84.33	43.64	40.68	21.16	27.37