Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation

Supplementary Material

1. More Implementation Details

The CLIP model with ViT-B/16 [3] is used as ExCEL's encoder, which is frozen during the training. For TSE module, we generate 20 descriptions from GPT-4 for each foreground category. We generate descriptions using OpenAI API, which can be seamlessly plugged into WSSS process. We save the descriptions beforehand to minimize redundant API calls. In our experiments, generating descriptions for VOC takes less than 4 minutes with no GPU usage, barely introducing extra burden for WSSS training. For the global template E_c, we adopt 'a clean origami [CLASS]' following previous methods [8, 16]. The number of attribute embeddings B is set to 112 and 224 for PASCAL VOC and MS COCO, respectively. The SVC module is conducted in the last $N = 5$ layers. The adapter in LVC module consists of 12 MLP layers and 1 convolutional layer. Our decoder adopts a simple Transformer-based head following [16], which contains 3 Transformer layers. Features F_l from each layer of CLIP are sent to it for the segmentation predictions. The weight factor λ of attribution information and w_i to balance attention map from $\{q, k, v\}$ are set as 1.0, and $1/3$ by default. The TOPK operation with a ratio of 0.9 is adopted to filter irrelevant attributes. The scaling and shifting factors, α and β are set as 3.0 and 1.0, respectively. The loss weight γ is set as 0.1. We conduct hyper-parameter tuning on VOC and directly transfer the parameter settings to COCO without extensive parameter optimization.

Following previous methods [10, 12, 15, 16], the training images are augmented with random horizontal flipping, random scaling with a ratio of $[0.5, 2.0]$, and random cropping into 320 × 320. The AdamW optimizer is used for training the adapter and decoder with a polynomial schedule. The learning rate is $1e-4$ and the weight decay is $1e-2$. The warm-up iterations are set as 50 and the warm-up learning rate is $1e-6$. The training iteration is set as $30,000$ for PASCAL VOC and $100,000$ for MS COCO. During the inference, the multi-scale and DenseCRF [4] post-processing techniques are used to refine the segmentation results. All experiments are conducted on a single RTX 3090 GPU.

2. More Quantitative Results

2.1. Analysis of Hyper-parameter

Attribute Weight. We introduce a weighting factor λ to balance the contribution of attribute knowledge. As shown in Tab. 1 (a), ExCEL maintains favorable performance when λ is set to 2.0 or 0.5 while the performance drops when we set it to a low value of 0.1 , which highlights the importance of implicit attributes to enrich the text representations. The best performance is achieved at $\lambda = 1.0$.

TOPK Filtering Ratio. We propose a TOP-K filtering operation to construct relevant attribute neighbors A_c and remove irrelevant ones. A filtering ratio K_r controls the number of neighbors. As shown in Tab. 1 (b), segmentation performance drops to 74.8% mIoU when $K_r = 0.1$, likely because excluding most useful attributes limits semantic enrichment for text prompting. It reports that ExCEL achieves the best performance when K_r is set to 0.9.

Number of SVC Layers. Unlike MaskCLIP [17] which uses v from the last CLIP layer, our SVC module mines fine-grained knowledge from intermediate layers. As shown in Tab. 1 (c), performance improves with more SVC layers, validating the effectiveness of this strategy. ExCEL achieves optimal results when SVC is applied to the last 5 layers of CLIP.

Scaling and Shifting Factors. We introduce scaling and shifting factors α and β to adjust dynamic correlations. Tab. 1 (d, e) analyzes their impact. ExCEL maintains consistent performance with varying α, while β plays a key role in filtering irrelevant correlations. When β is set to 2.0, the performance drops significantly to 65.3% mIoU, likely due to excessive suppression of correlations with similar semantics, hindering dense knowledge extraction. ExCEL achieves optimal performance when β is set to 1.0.

Diversity Loss Weight. Tab. 1 (f) evaluates the impact of the loss weight γ, which balances the contribution of the diversity loss $\mathcal{L}_{\text{div}}$. When $\gamma = 0.1$, ExCEL achieves optimal performance, indicating an effective trade-off between diversity and segmentation losses.

2.2. Types of Text Prompting

In Tab. 2 (a), we compare different text prompting strategies: the global template (Global), fusing descriptive embeddings into the final text embedding (Fuse), and our implicit attribute-hunting approach. The results demonstrate that treating text prompting as an implicit attribute-hunting process outperforms other methods. Qualitative comparisons in Fig. 6 (g-i) further illustrate that our method produces more precise and complete CAMs.

2.3. Analysis of Intra-correlation

Suboptimal q-k Attention in CLIP. CLIP's visual features lack fine-grained details, resulting in unreasonable ob-

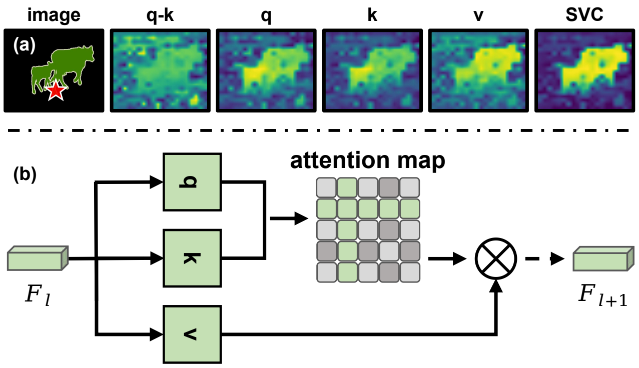

Figure 1. Illustration of how the original q-k attention homogenizes diverse tokens from $\{q, k, v\}$, resulting in limited diversity in attention maps. (a) Visualization of attention features from q-k, q, k, v, and our SVC. The red star indicates the query patch. (b) Illustration of the original q-k attention mechanism.

ject activations. We attribute this to suboptimal q-k attention, which homogenizes the diverse tokens in $\{q, k, v\}$. The q-k attention mechanism as shown in Fig. 1 (b). To further explore this, we visualize the q-k and $\{q, k, v\}$ attention maps in Fig. 1 (a). The results reveal that $\{q, k, v\}$ attention preserves essential spatial details, while q-k attention loses diversity, supporting our claim. Based on these findings, we perform Intra-correlation within each space of $\{q, k, v\}$, thereby avoiding the smoothing effects of q-k attention, significantly enriching visual features with spatial details and generating better attention maps than using $\{q, k, v\}$ alone, as demonstrated in Fig. 1 (a) and Fig. 8.

Types of Intra-correlation. Tab. 2 (b) examines different types of Intra-correlation. Compared to applying individual intra-correlation within each of $\{q, k, v\}$ separately, our approach, which combines attention maps from all three spaces, demonstrates consistent superiority.

2.4. Training Convergence

Fig. 2 shows the training convergence of ExCEL compared with CLIP-based SOTA WeCLIP [16]. We evaluate the performance on PASCAL VOC val set and post-processing methods, such as DenseCRF [4], are not used during the training. It can be seen that our ExCEL consistently outperforms WeCLIP throughout the training process.

2.5. Category-wise Performance

We report detailed confusion ratio (CR) performance on the VOC val set and compare ExCEL with recent methods in Tab. 3. CR is calculated by FP/TP, the lower the better [15]. It shows that ExCEL consistently achieves a lower CR among 10 categories and generates 0.20 mCR for the average, which demonstrate the superiority of our dense patch-text alignment in generating more precise predictions.

In Tab. 4, the per-category comparisons between our method ExCEL and other recent methods are conducted on PASCAL VOC val set. For recent CLIP-based SOTA WeCLIP, our method consistently demonstrates better IoU in all categories and significantly outperforms it by 2.0% mIoU. Compared to other multi-stage methods, such as MCTformer+ [14], our approach holds significant superiority with 4.4% mIoU as well. It is noted that our method is trained in a single-stage paradigm and only the adapter and segmentation head are optimized during the training process. It reveals the great potential of our designed patch-text alignment paradigm for efficient WSSS.

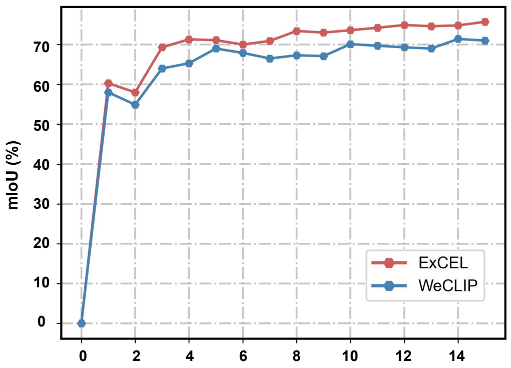

Figure 2. Training convergence comparison between ExCEL and WeCLIP on VOC val set. CRF is not used during the training. The result of WeCLIP is reproduced using the official code.

3. More Qualitative Results

3.1. Visualization of CAM

Fig. 3 presents additional qualitative ablations of our modules alongside comparisons with recent CLIP-based methods [8, 16, 17]. The results clearly validate the effectiveness of our proposed components and highlight the superiority of the patch-text alignment paradigm to generate more precise and complete CAMs over image-text alignment approaches.

3.2. Visualization of Segmentation.

Additional segmentation results on the PASCAL VOC and MS COCO val sets are presented in Fig. 4 and Fig. 5, respectively. Leveraging CLIP's dense knowledge through the proposed patch-text alignment, our method delivers more complete and precise predictions with sharper boundaries compared to recent approaches [15, 16].

3.3. Visualization of Attribute Response

Fig. 6 (b-f) shows more attribute responses from our TSE module. 5 implicit attributes are sampled. The clustered attributes effectively capture distinct parts of objects, validating that our attribute hunting process enriches text semantics by aggregating related features and enabling more comprehensive visual responses. Additionally, Fig. 6 (g-i) compares different text prompting methods: ”Fuse,” which directly combines n descriptive embeddings per class, and "Global," which uses the template "a clean origami [CLASS]." The results highlight the superiority of our implicit attribute hunting process in producing more complete CAMs while minimizing irrelevant noise.

3.4. Visualization of Feature Representation

Fig. 7 and Fig. 8 provide additional visual comparisons of feature representations. As shown in Fig. 7, given the query patch marked by a red star, the proposed SVC generates attention maps with clearer boundaries compared to the CLIP baseline and MaskCLIP. It also produces more diverse pairwise token relations, where semantically related token pairs exhibit higher similarities. By incorporating the optimized LVC, our method further enhances CLIP's dense capabilities, yielding features with richer spatial details.

Fig. 8 illustrates how the original q-k attention homogenizes diverse tokens from $\{q, k, v\}$, resulting in limited diversity in attention maps. This observation motivates us to perform Intra-correlation within each space of $\{q, k, v\}$, thereby avoiding the smoothing effects of q-k attention. While $\{q, k, v\}$ individually capture more diverse details, they remain insufficient to precisely highlight foregrounds. Our SVC effectively integrates attention maps from each space and mines fine-grained knowledge from intermediate layers, consistently generating more diverse and precise attention maps than $\{q, k, v\}$ alone. Additionally, incorporating the LVC module further enhances performance.

References

[1] Nikita Ara. and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020. 4

[2] Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, and Lei Zhang. Fpr: False positive rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1108–1118, 2023. 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011. 1, 2, 4

[5] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. 4

[6] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 4

[7] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021. 4

[8] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *arXiv preprint arXiv:2212.09506*, 2022. 1, 2, 5

[9] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022. 4

[10] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2303.01267*, 2023. 1, 4

[11] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021. 4

[12] Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3534–3543, 2024. 1, 4

[13] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 4

[14] Lian Xu, Mohammed Bennamoun, Farid Boussaid, Hamid Laga, Wanli Ouyang, and Dan Xu. Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 2, 4

[15] Zhiwei Yang, Kexue Fu, Minghong Duan, Linhao Qu, Shuo Wang, and Zhijian Song. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3615, 2024. 1, 2, 4, 5, 6

[16] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3796–3806, 2024. 1, 2, 4, 5, 6

[17] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2, 5, 7

Table 1. Impact of hyper-parameters on PASCAL VOC. M: quality of pseudo mask on the train set. Seg.: semantic prediction performance on val set. Attr.: Attribute. The post-processing techniques, such as DenseCRF [4], are not adopted.

(a) Attribute Weight λ.

Attr. weight λ	M	Seg.
0.1	76.6	75.7
0.5	77.3	76.5
1.0	**78.0**	**77.2**
2.0	77.4	76.7

(b) TOPK ratio K_r.

TOPK ratio K_r	M	Seg.
0.1	75.6	74.8
0.5	76.5	75.8
0.9	**78.0**	**77.2**
1.0	77.8	76.9

(c) Number of SVC layers N.

SVC layers N	M	Seg.
1	71.9	71.1
3	76.3	75.4
5	**78.0**	**77.2**
8	76.8	75.9

(d) Scaling Factor α.

Scaling factor α	M	Seg.
1.0	77.5	76.6
2.0	77.6	76.4
3.0	**78.0**	**77.2**
5.0	77.7	76.9

(e) Shifting Factor β.

Shifting factor β	M	Seg.
0.5	77.6	76.9
1.0	**78.0**	**77.2**
1.5	75.7	74.7
2.0	67.6	65.3

(f) Loss Weight γ.

Loss weight γ	M	Seg.
0.1	**78.0**	**77.2**
0.3	77.2	76.4
0.5	76.9	75.9
1.0	76.0	75.1

Table 2. Different types of text prompting and Intra-correlation. Global means the global template of each class, i.e., ”a clean origami of [CLASS]”. Fuse means descriptive embeddings are directly fused into the final text embedding. The post-processing, such as DenseCRF, is not adopted. The experiments are conducted on PASCAL VOC val set.

(a) Types of Text prompting.

Types	Global	Fuse	Ours
Precision	83.6	83.8	85.0
Recall	86.9	87.3	88.4
Seg.	74.9	75.1	77.2

(b) Types of Intra-correlation.

Types	q-k	q-q	k-k	v-v	Ours
Precision	19.2	84.4	82.9	84.6	85.0
Recall	21.5	86.6	82.9	85.2	88.4
Seg.	12.0	75.1	75.6	75.0	77.2

Table 3. Per-category confusion ratio comparison with recent methods on PASCAL VOC val set. Confusion ratio (CR) [15] is calculated by FP/TP, the lower the better. †: Our reproduction following the official code.

Methods	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mCR
AFA [9] CVPR'2022	0.05	0.12	2.16	0.14	0.42	0.20	0.10	**0.06**	0.20	1.09	0.07	0.16	0.12	0.13	0.27	0.34	0.39	0.09	0.57	0.63	0.49	0.36
ToCo [10] CVPR'2023	0.04	0.19	**0.84**	0.42	1.11	0.13	0.11	0.11	**0.02**	0.65	0.03	0.32	0.08	0.09	0.21	0.06	0.59	0.06	0.77	0.75	0.34	0.32
DuPL [12] CVPR'2024	0.03	0.26	1.00	0.20	0.53	0.17	0.10	0.20	0.03	0.71	0.02	0.23	0.08	0.05	**0.16**	0.10	0.54	0.04	0.59	0.47	1.10	0.31
SeCo [15] CVPR'2024	0.04	0.07	1.22	0.10	0.32	0.17	**0.09**	0.07	0.02	**0.48**	**0.02**	0.28	0.09	**0.05**	0.17	**0.06**	**0.29**	0.04	0.35	0.54	0.45	0.23
†WeCLIP [16] CVPR'2024	0.03	0.09	1.29	0.07	0.19	0.15	0.10	0.12	0.05	0.79	0.05	0.17	0.05	0.11	0.21	0.10	0.44	0.07	0.44	0.10	**0.27**	0.23
ExCEL(Ours)	**0.03**	**0.06**	1.14	**0.04**	**0.13**	**0.13**	0.10	0.11	0.04	0.69	0.03	**0.12**	**0.04**	0.08	0.20	0.08	0.36	**0.03**	**0.34**	**0.09**	0.32	**0.20**

Table 4. Per-category performance comparison with recent methods on PASCAL VOC val set. IoU is the metric.

Methods	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
Multi-staged methods																						
CDA [2] CVPR'2021	89.1	69.7	34.5	86.4	41.3	69.2	81.3	79.5	82.1	31.1	8.3	50.8	80.6	76.1	72.2	77.6	48.8	81.2	42.5	60.6	54.3	66.1
AdvCAM [5]CVPR'21	89.5	76.9	33.5	80.3	63.7	68.6	**89.7**	77.9	87.6	31.6	77.2	36.2	82.6	78.7	73.5	69.8	51.9	81.9	43.8	70.9	52.6	67.5
EPS [7]CVPR'21	91.7	89.4	40.6	84.7	67.0	71.6	87.8	82.7	87.4	33.6	81.9	37.3	82.5	82.9	76.6	82.8	54	79.7	39.1	85.4	51.7	71.0
W-OoD [6]CVPR'22	91.0	80.1	34.1	88.1	64.8	68.3	87.4	84.4	89.8	30.1	87.8	34.7	87.5	85.9	79.8	75.0	56.4	84.5	47.8	80.4	46.4	70.7
FPR [11]ICCV'23	91.4	81.8	35.1	82.4	68.7	73.7	88.8	80.5	85.9	33.3	82.4	45.3	82.5	81.6	72.9	78.5	50.7	82.6	46.5	83.1	49.1	70.3
MCTformer [13]CVPR'22	91.9	78.3	39.5	89.9	55.9	76.7	81.8	79	90.7	32.6	87.1	57.2	87	84.6	77.4	79.2	55.1	89.2	47.2	70.4	58.8	71.9
MCTformer+ [14]TPAMI'24	93.3	87.0	37.8	91.1	66.8	**79.9**	87.4	82.2	91.3	32.1	84.8	58.8	86.2	82.2	79.0	82.2	54.4	87.5	50.0	82.0	57.3	74.0
Single-staged methods																						
1Stage [1]CVPR'20	88.7	70.4	35.1	75.7	51.9	65.8	71.9	64.2	81.1	30.8	73.3	28.1	81.6	69.1	62.6	74.8	48.6	71.0	40.1	68.5	64.3	62.7
AFA [9]CVPR'22	89.7	79.3	30.3	79.8	64.6	62.0	82.3	66.5	80.5	29.6	83.9	45.0	80.2	76.0	70.1	76.1	51.8	84.8	44.6	59.6	52.8	66.0
ToCo [10]CVPR'23	91.1	80.6	**48.6**	68.4	45.4	79.7	87.3	83.3	89.9	35.7	84.7	**60.5**	83.7	83.4	76.7	83.0	56.5	88.0	43.8	60.4	63.1	71.1
DuPL [12]CVPR'24	91.8	77.9	47.0	81.7	58.7	78.4	88.8	77.5	91.9	38.1	91.5	55.5	87.9	**90.0**	77.7	**85.9**	60.7	**92.7**	53.9	66.1	45.5	73.3
SeCo [15]CVPR'24	92.5	86.3	39.8	88.8	68.4	78.5	88.1	80.1	90.4	38.3	84.5	52.4	86.9	85.9	73.5	84.4	62.4	89.6	57.4	62.2	62.6	74.0
†WeCLIP [16]CVPR'24	93.5	87.8	41.1	90.6	74.4	69.2	88.4	84.2	91.7	41.0	90.7	57.1	89.7	86.1	79.5	82.0	59.7	84.0	59.7	84.0	65.3	76.4
ExCEL(Ours)	**94.1**	**90.2**	43.1	**91.8**	**77.3**	72.6	88.4	**84.6**	**93.4**	**44.5**	**91.6**	59.3	**90.8**	87.6	**80.2**	83.0	**65.7**	92.2	**64.1**	**84.6**	**66.4**	**78.4**

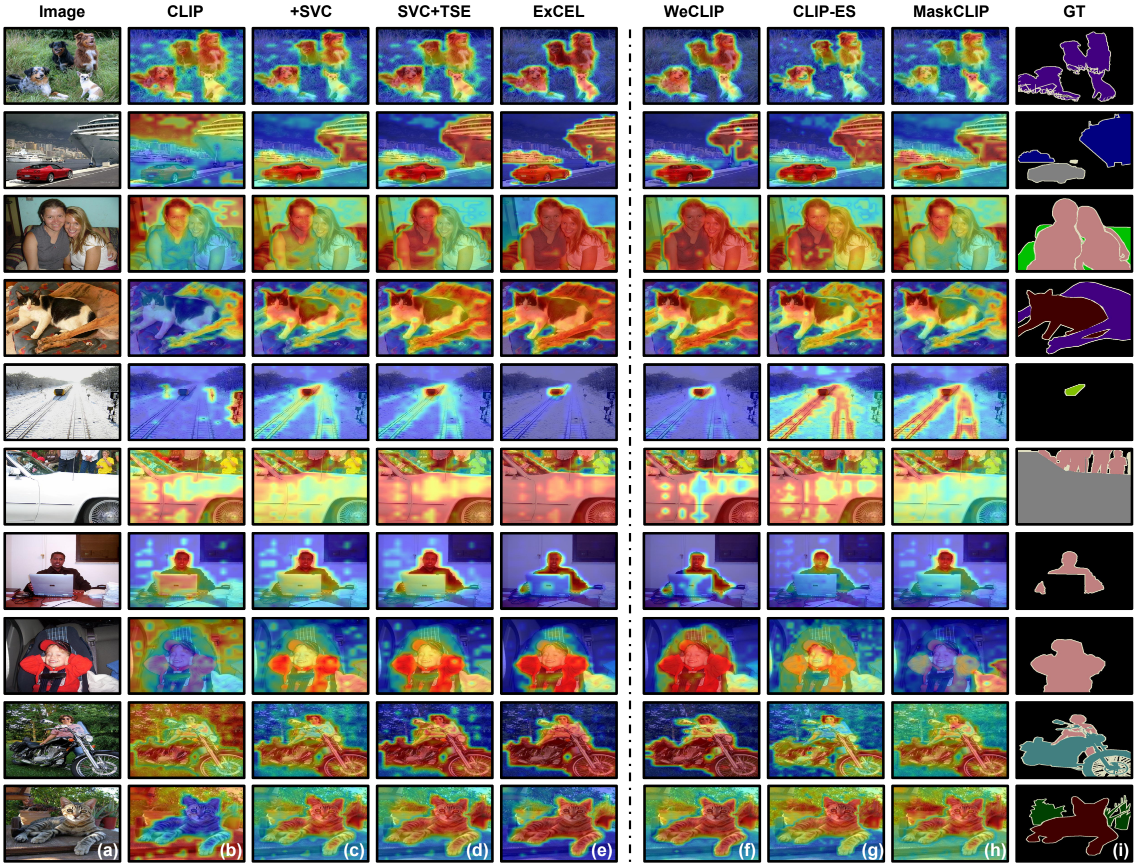

Figure 3. CAM visualizations on PASCAL VOC train set. (a) Image. (b-e) Qualitative ablations of our key components. (e-h) CAM comparisons between our ExCEL and recent methods, i.e., WeCLIP [16], CLIP-ES [8], and MaskCLIP [17]. (i) Ground truth.

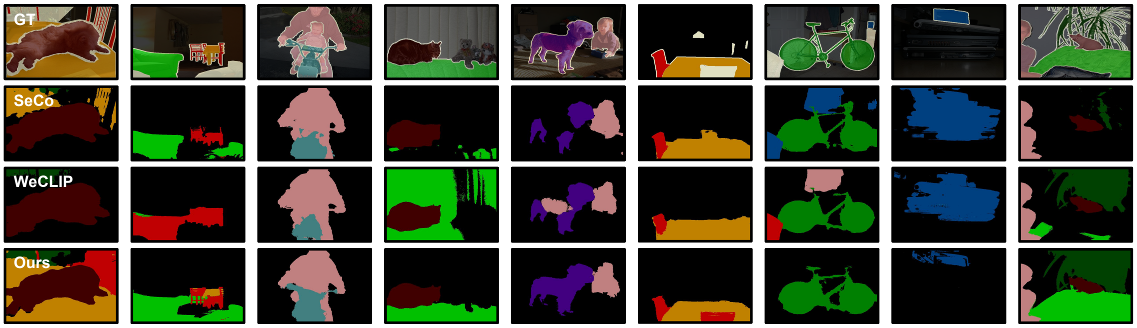

Figure 4. Qualitative segmentation performance on PASCAL VOC. The comparisons are conducted among SeCo [15], WeCLIP [16], and ours. Our ExCEL produces more precise segmentation predictions.

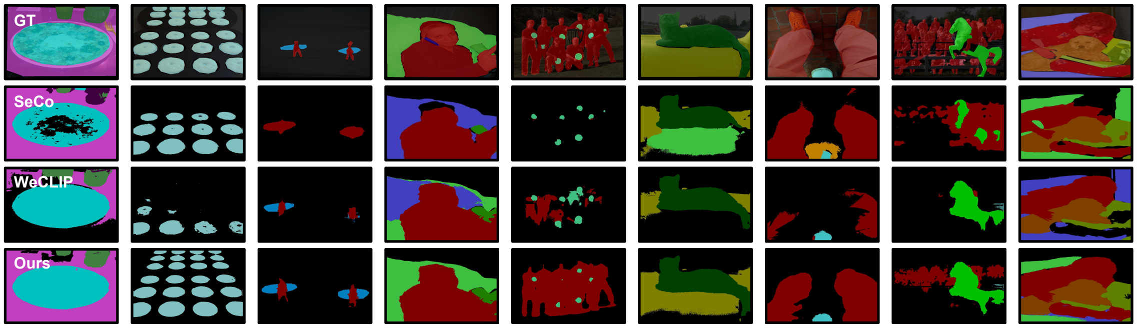

Figure 5. Qualitative segmentation performance on MS COCO. The comparisons are conducted among SeCo [15], WeCLIP [16], and ours. ExCEL generates better predictions than other methods.

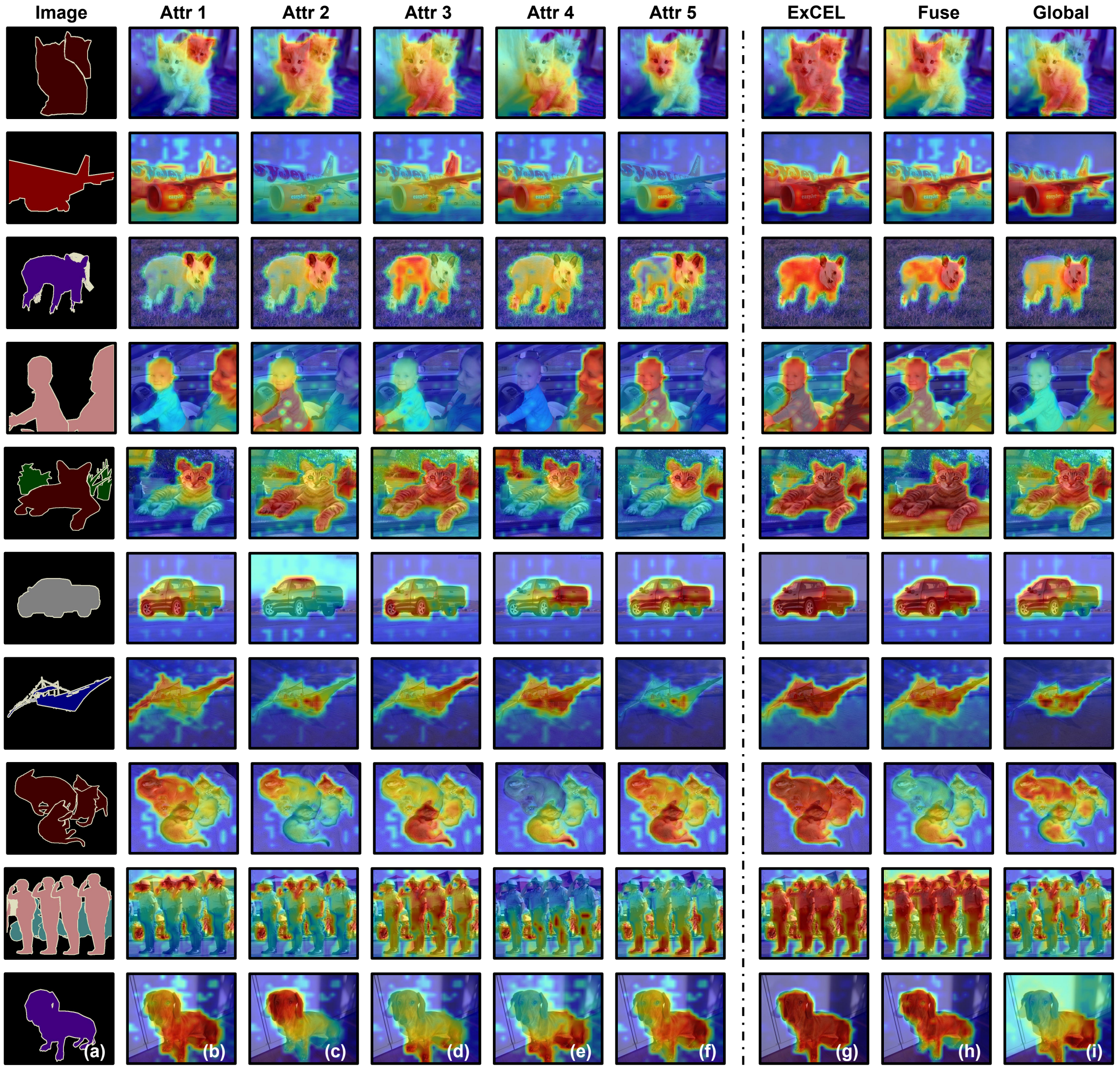

Figure 6. Implicit attribute responses. (a) Ground truth. (b-f) 5 attributes are sampled to draw the visualizations. They highlight different parts of objects. (g-i) Three types of text prompting are also visualized, i.e., our attribute hunting operation, Fuse," which directly combines n descriptive embeddings per class, and "Global," which uses the template "a clean origami [CLASS]." Our method shows advantages in generating more complete and precise object activations.

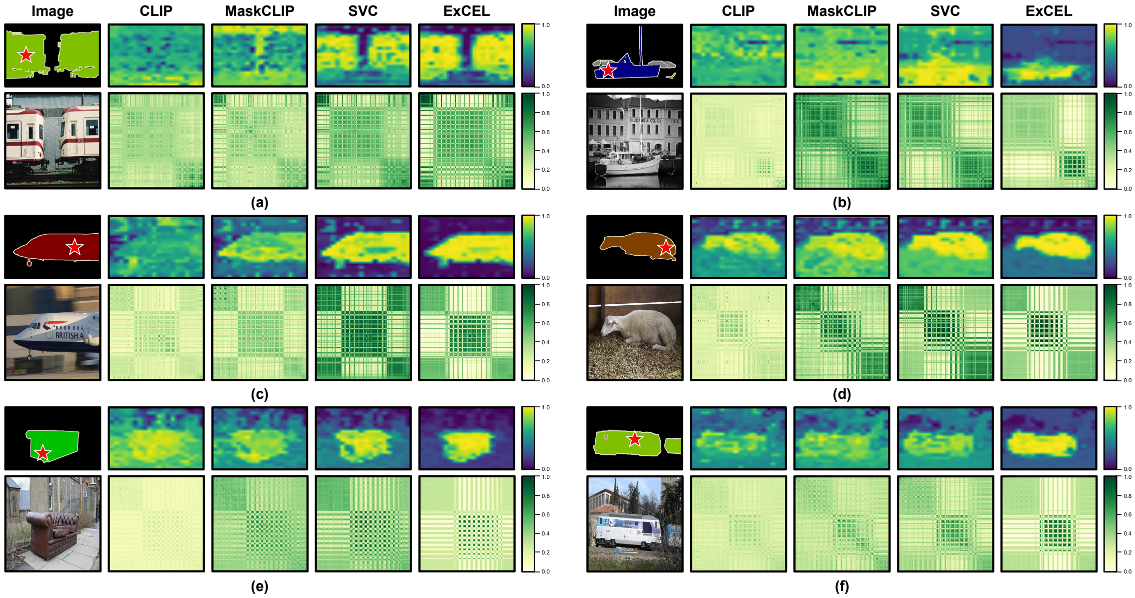

Figure 7. Visualization of attention features from the last visual layer of CLIP. Given the query patch (marked by a red star in the first row of each case), our SVC module and optimized ExCEL generate more diverse attention maps with fine-grained spatial information compared to CLIP's q-k attention or MaskCLIP's v-v attention [17]. The second row in each case indicates the pair wise token relations. It shows that our approach distinctly groups tokens with similar semantics, aligning pairwise similarities with corresponding semantics.

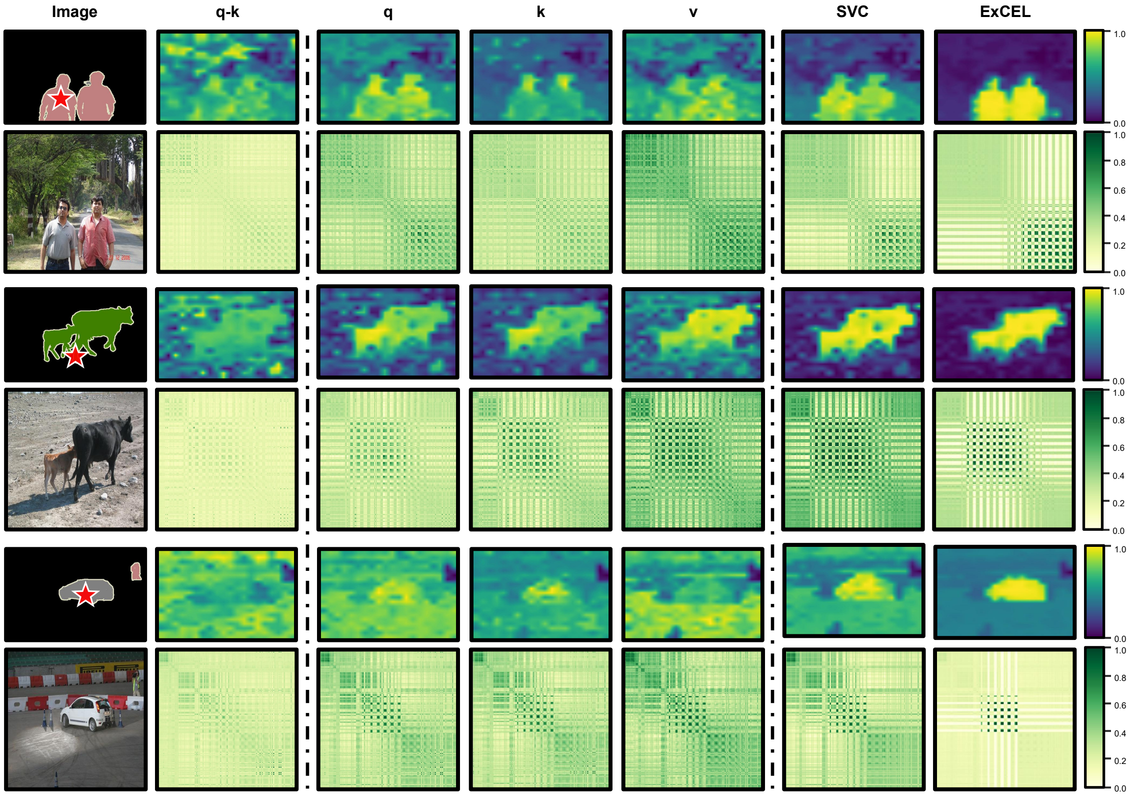

Figure 8. Visualization of q-k, q, k, v, and our attention maps given the query patch. It shows that the original q-k attention homogenizes the diverse tokens from q, k, v and falls short in generating diverse attention maps. Our method conducts Intra-correlation within each space of q, k, v, avoiding the smoothing effect of q-k attention and generating more diverse attention features.