$\mathcal{DIH}$-CLIP: Unleashing the Diversity of Multi-Head Self-Attention for Training-Free Open-Vocabulary Semantic Segmentation

Supplementary Material

Table 1. Ablation studies of mid-layer features.

Layers	VOC21	Context60	VOC20	Stuff	City
2-11	62.1	34.2	83.9	25.1	37.6
3-10	64.2	36.0	84.9	26.7	40.2
4-9	63.6	35.9	84.7	26.4	39.9
5-8	63.4	35.6	84.2	26.1	39.5

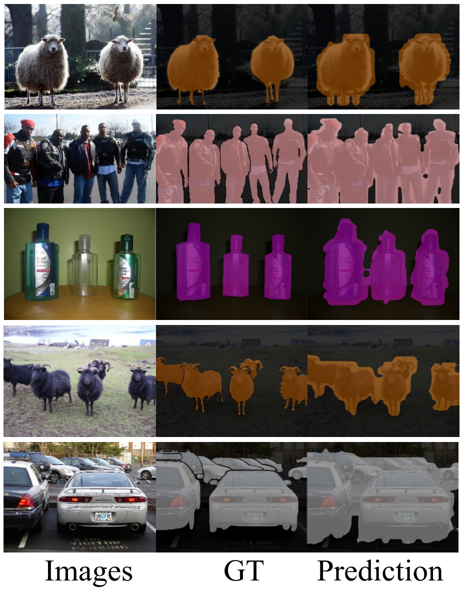

Figure 1. The semantic segmentation results of $\mathcal{DIH}$-CLIP on multi-object scene.

1. The hyper-parameter of middle layer

To identify the best selection of middle-layer features, we use SHA to fuse features from different middle layers. The comparison results are reported in Table 1. Results indicate that the front and back layers of the CLIP visual encoder can provide limited detail, while the middle layer features can often provide effective detail for train-free open-vocabulary semantic segmentation. Of cause, the features from 3-10 layers are most effective.

2. Visualization

We further visualize the segmentation results of our $\mathcal{DIH}$-CLIP. The qualitative zero-shot segmentation results shown in Fig. 1 and Fig. 2 demonstrate that our $\mathcal{DIH}$-CLIP model yields clear segmentation masks in most cases.

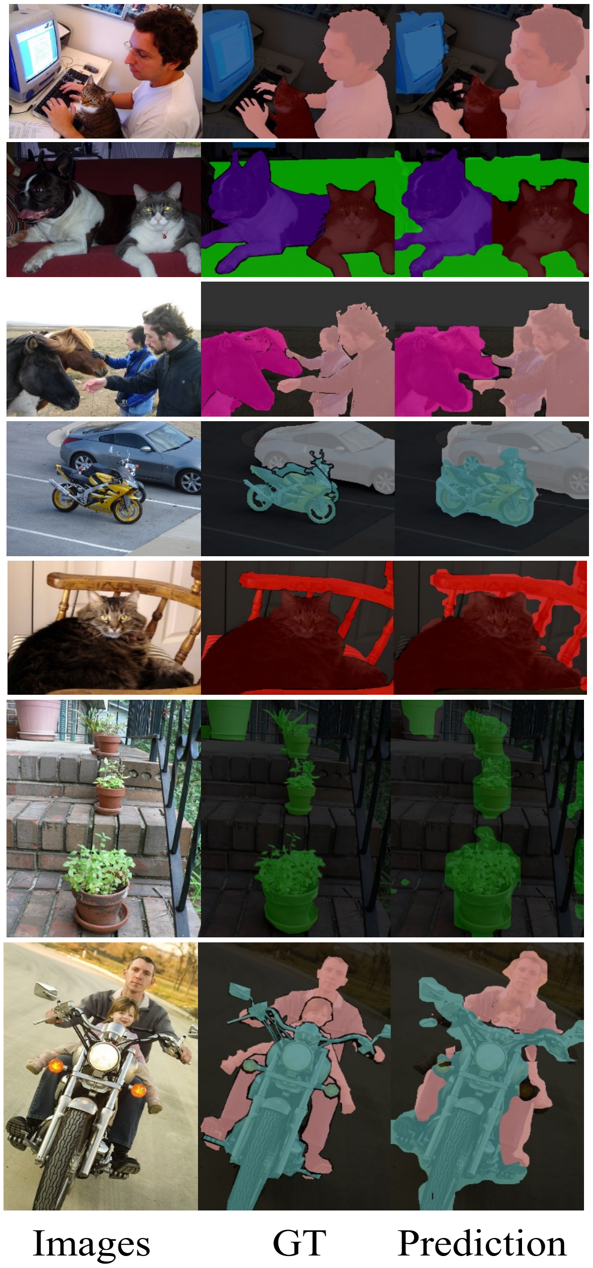

Figure 2. The semantic segmentation results of $\mathcal{DIH}$-CLIP on multi-label scene.

3. Extreme case of SHA

To assess the robustness of our method in extreme cases. We evaluate performances of rare classes on Pascal Con-

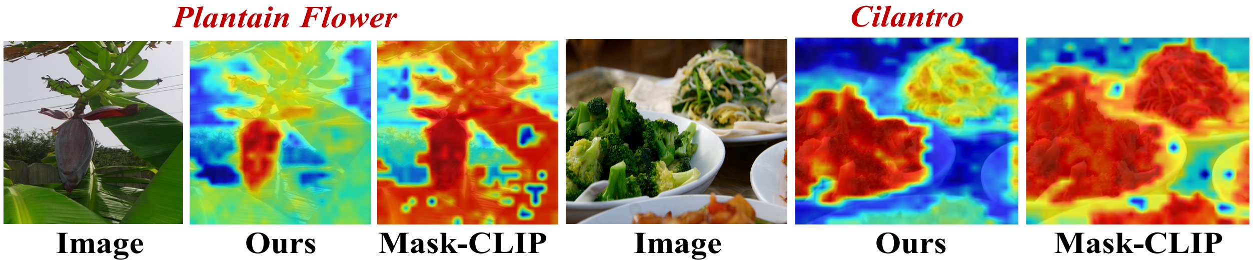

Figure 3. Visualization on the samples with novel texts.

Table 2. Performance on rare classes of pascalcontext with mIoU.

Method	shelves	light	ceiling	sidewalk	track
SCLIP	8.1	14.1	30.2	8.3	12.7
Ours	9.1	14.4	31.0	9.7	13.5

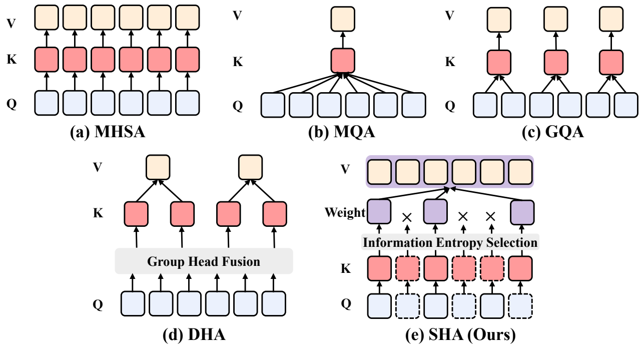

Figure 4. Different attention comparisons.

text59 in Table 3, where our $\mathcal{DIH}$-CLIP still significantly outperforms SCLIP. The results prove the robustness of our method.

4. Potential information loss of SHA

The proposed SHA deletes some head attention, which raises concerns about potential information loss. Here, we declare that $\mathcal{DIH}$-CLIP doesn't cause potential information loss for two reasons: (1) SHA retains about 42% $\backsim$ 75% attention maps of MHSA; (2) although SHA deletes redundant attention maps, we employ all V features to generate visual features. Besides, to prove the effectiveness and robustness of our method in rare scenarios, we present some complex samples with novel text prompts in Fig. 3, which proves that $\mathcal{DIH}$-CLIP is effective for complex and unseen scenes. In particular, our $\mathcal{DIH}$-CLIP precisely identifies fine-grained categories, such as cilantro in Figure 3, while Mask-CLIP recognizes cilantro as a coarse-grained vegetable.

5. SHA *vs* other attentions

To reduce the high-cost reasoning of the self-attention mechanism, researchers have proposed some attention substitution methods based on pruning ideas, as shown in Fig. 4. FRMA [3] manually selects the best individual head through mIoU scores of the GT evaluation. The MHSA [5], MQA [4], GQA [1], and DHA [2] all adopt shared key/value features to match a group of query features for generating the attention maps, which are proposed to accelerate the reasoning of large language models. The differences between our SHA and theirs are shown in Fig. 4. The comparison results indicate that our SHA is efficient for CLIP-based TF-OVSS.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. 3

[2] Yilong Chen, Linhao Zhang, Junyuan Shang, Zhenyu Zhang, Tingwen Liu, Shuohuan Wang, and Yu Sun. Dha: Learning decoupled-head attention from transformer checkpoints via adaptive heads fusion. *Advances in Neural Information Processing Systems*, 37:45879–45913, 2024. 3

[3] Dong Un Kang and Se Young Chun. Focusing on representation of multi-head attention for open-vocabulary semantic segmentation. In *2025 International Conference on Electronics, Information, and Communication*, pages 1–3, 2025. 3

[4] Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019. 3

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances In Neural Information Processing Systems*, 30, 2017. 3