CorrCLIP: Reconstructing Patch Correlations in CLIP for Open-Vocabulary Semantic Segmentation

Supplementary Material

Intra-Class and Inter-Class Correlations. We explore the impact of inter- and intra-class correlations on CLIP's segmentation capabilities on other datasets, as shown in Fig. 1. Consistent with the findings in the main text, intra-class correlations can significantly enhance CLIP's segmentation performance, while inter-class correlations impair its segmentation capability.

Performance on Out-of-Distribution Datasets. Fully-supervised methods [4, 10] are typically trained on the COCO dataset, which shares high similarity with the five datasets used in the main text. This setup, however, does not adequately showcase the full potential of OVSS. Therefore, we conduct additional evaluations on a selection of datasets from the MESS benchmark [2] that deviate from the COCO distribution. As shown in Tab. 1, CorrCLIP outperforms the fully-supervised method on these datasets, demonstrating that it effectively leverages CLIP's open-vocabulary capabilities.

Discussion on Spatial Branch. In ViT, patch features are progressively transformed through successive layers, which critically employ residual connections. This design ensures that features from lower layers are directly added to higher-level ones, thereby retaining a degree of similarity to the features in the final layer and serving as a complement to the final patch features. As demonstrated in Tab. 2, the respectable segmentation performance achieved using only the Spatial Branch corroborates this hypothesis. However, the results on the COCO dataset, indicate that feature misalignment can still lead to performance degradation. Thus, there is still scope for refining this approach, for example, by selectively selecting the most suitable lower layers' features.

Computational Analysis. While the main body of this work focused on enhancing the CLIP model's segmentation capabilities, computational efficiency was not a primary design constraint. We now turn to an analysis of the computational cost associated with each component of our method and suggest avenues for its improvement.

The primary computational burden stems from the number of SAM sampling points as shown in Tab. 3. Reducing uniform sampling from 32×32 to 8×8 substantially improves processing speed with only marginal performance degradation, indicating that many sampled points are repetitive and optimizing sampling strategies is a key direction for acceleration. Another critical avenue for speed improvement lies in the adoption of a more efficient mask generator.

The computational cost of Map Correction is due to the cyclic updating of categories in the mask. Designing a parallel algorithm can further improve its speed.

Method	FoodSeg103 [14]	ATLANTIS [5]	CUB-200 [13]	SUIM [6]
OVSeg-L [10]	16.4	33.4	14.0	38.2
CAT-Seg-L [4]	30.5	33.6	9.2	54.0
CorrCLIP-L	**36.5** +6.0	**40.1** +6.5	**31.3** +17.3	**58.0** +4.0

Table 1. Comparison with fully-supervised methods on some datasets in MESS.

Branch	VOC20	PC59	Stuff	ADE	City
Main	88.7	46.2	30.6	25.3	48.3
Spatial	74.5	43.8	27.7	24.1	46.3

Table 2. Segmentation performance achieved using only the Spatial Branch.

The computational overhead of Feature Refinement arises from the addition of extra mask class tokens. The computational burden of Value Reconstruction stems from incorporating the DINO model, which can be mitigated by adopting smaller model variants (e.g., "DINO-S") to enhance speed and reduce memory footprint.

The clustering algorithm employed in Mask Merging is currently implemented on the CPU, with the potential for future acceleration through the use of GPU.

To enhance the real-world applicability of CorrCLIP, we implement two modifications to mitigate its high computational cost. First, we substitute SAM with more efficient mask generators. Second, we streamline CorrCLIP by removing mask merging and value reconstruction, resulting in a speed increase with an acceptable performance loss. As a result, CorrCLIP now achieves inference speeds comparable to, or even exceeding, those of ProxyCLIP and Trident, while retaining superior performance. Note that EoMT [7] and EntitySeg [11] perform better because they are better suited for our method's whole-image segmentation needs than SAM, which is designed for promptable segmentation.

More Qualitative Comparisons. We provide more qualitative comparisons between our CorrCLIP and existing methods, including ClearCLIP [8], ProxyCLIP [9], SC-CLIP [1], and Trident [12] in Figs. 2 to 6. Through explicit interaction scope restriction, our approach successfully corrects misclassifications that persist in other methods. Additionally, CorrCLIP exhibits three notable advantages: superior object continuity preservation, effective noise suppression, and enhanced robustness in challenging scenar-

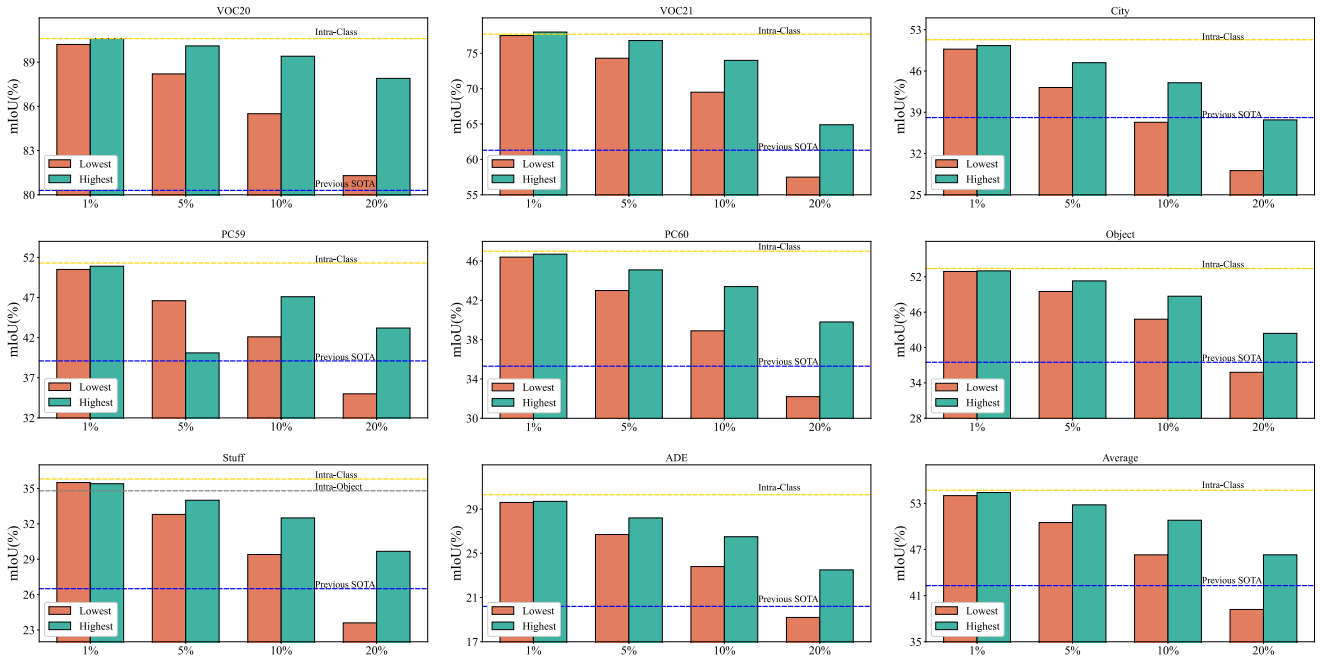

Figure 1. Impact of intra-class and inter-class correlations on CLIP's segmentation capability on the other seven benchmarks and the average of all benchmarks.

Method	Time(ms/image) ↓	Memory(MB) ↓	Parameter(M) ↓	Performance(mIoU) ↑
ClearCLIP [8]	19	718	150	38.1
ProxyCLIP [9]	69	1936	235	42.3
Trident [12]	81	2487	364	45.8
Sampled Points: 8×8				
+ Sope Reconsturction	116	2674	373	45.3
+ Map Correction	119	2674	373	47.5
+ Feature Refinement	132	2674	373	49.7
+ Value Reconstruction	170	2997	458	50.2
+ Mask Merging	177	2997	458	50.4
Sampled Points: 32×32				
+ Sope Reconstruction	1111	2902	373	46.8
+ Map Correction	1115	2902	373	49.2
+ Feature Refinement	1129	2902	373	50.5
+ Value Reconstruction	1168	3208	458	50.8
+ Mask Merging	1258	3208	458	51.0
Faster Version				
CorrCLIP$_{\text{M2F}}$	81	1765	366	49.2
CorrCLIP$_{\text{EoMT}}$	**56**	1818	466	**51.6**
CorrCLIP$_{\text{EntitySeg}}$	78	**1100**	**197**	51.1

Table 3. Computational costs on RTX 4090 with FP16. Performance is average mIoU across eight benchmarks. The blue subscript indicates the mask generator. M2F (Mask2Former [3]) uses a Swin-L backbone, and EoMT uses ViT-L; both are pretrained on COCO Panoptic. EntitySeg uses a Swin-T backbone.

ios where alternative methods often exhibit fragmentation or semantic inconsistency. These qualitative results corroborate our method's capability to establish more reliable visual-semantic correspondences.

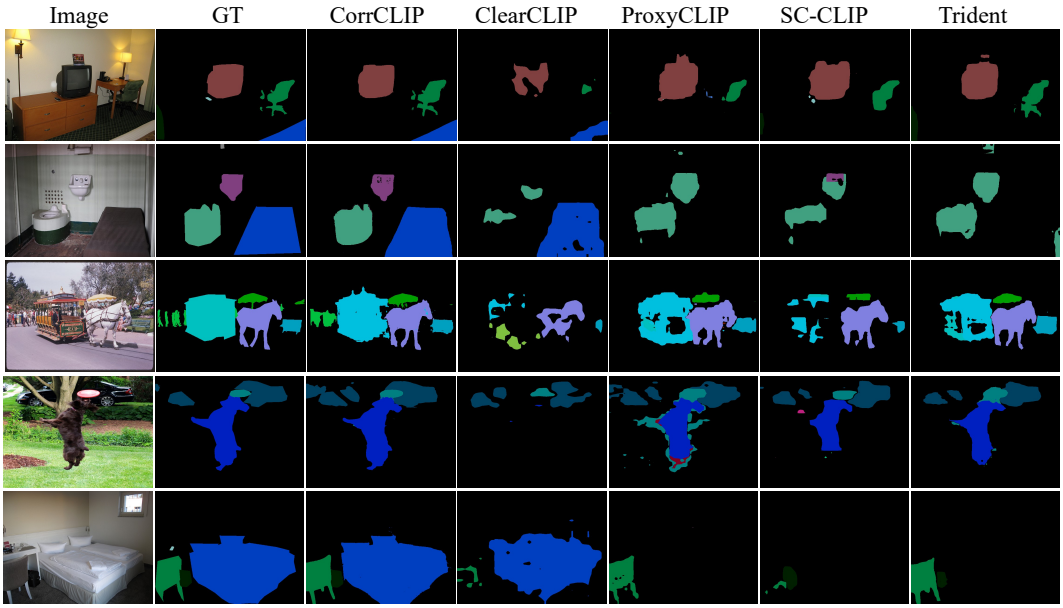

Figure 2. More qualitative comparison of segmentation maps between our method, CorrCLIP, and the other four methods on Object.

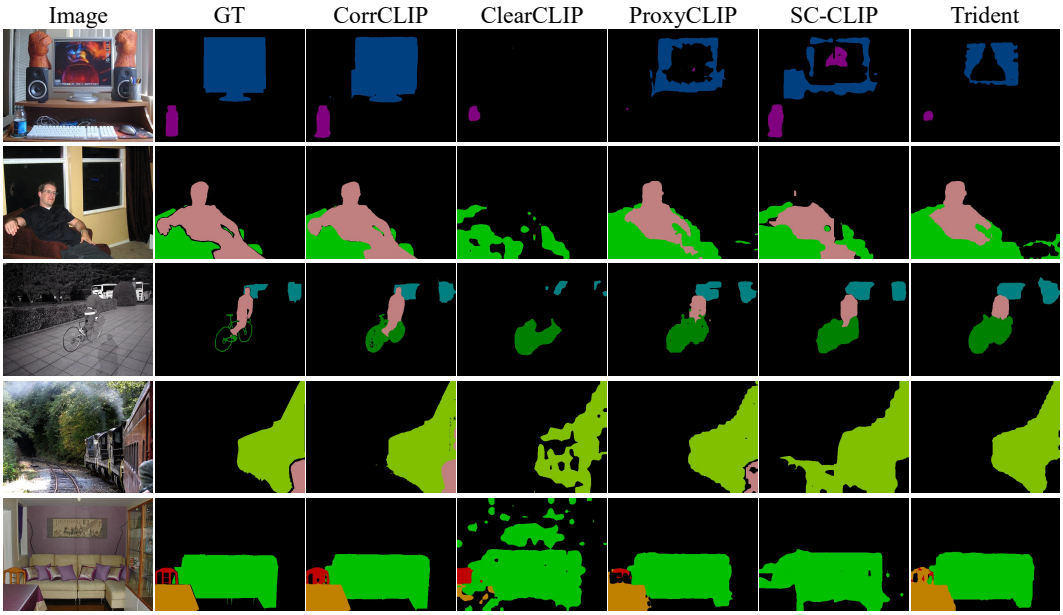

Figure 3. More qualitative comparison of segmentation maps between our method, CorrCLIP, and the other four methods on VOC21.

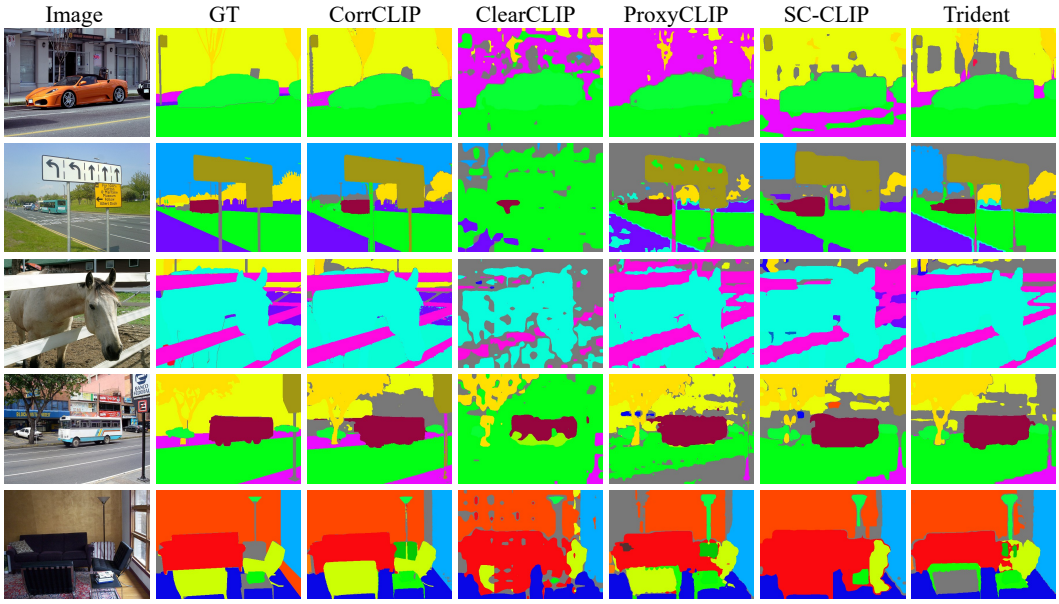

Figure 4. More qualitative comparison of segmentation maps between our method, CorrCLIP, and the other four methods on PC60.

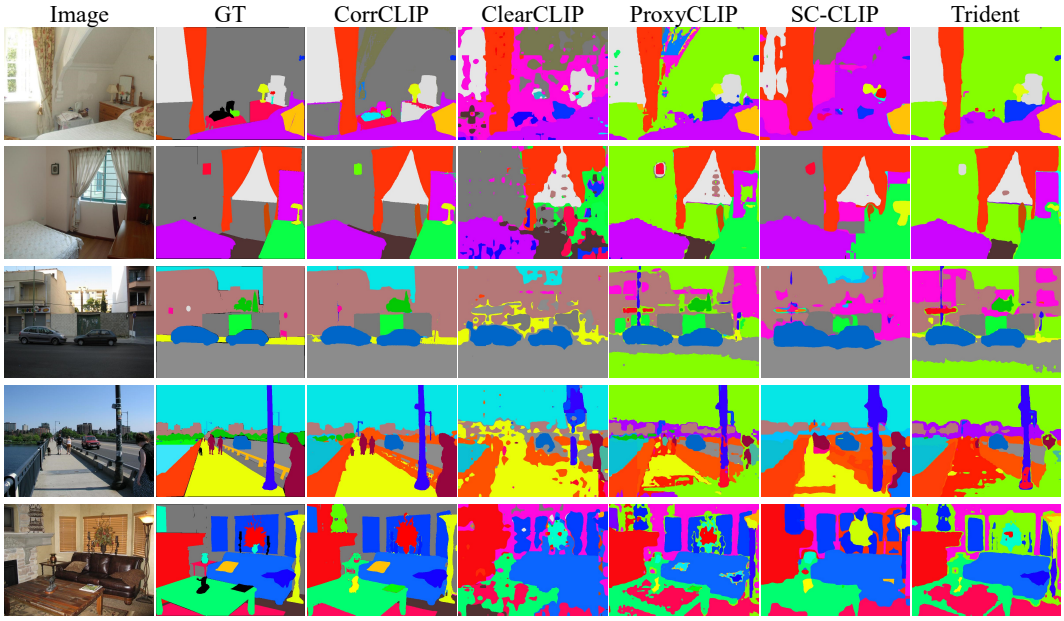

Figure 5. More qualitative comparison of segmentation maps between our method, CorrCLIP, and the other four methods on ADE.

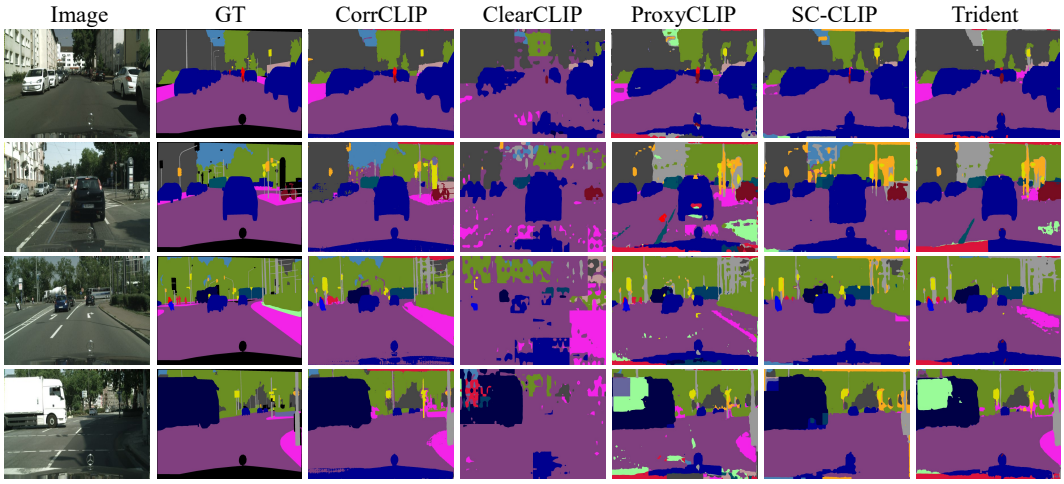

Figure 6. More qualitative comparison of segmentation maps between our method, CorrCLIP, and the other four methods on City.

References

[1] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 1

[2] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kühne, and Michael Vössing. What a mess: Multi-domain evaluation of zero-shot semantic segmentation. *NeurIPS*, 36:73299–73311, 2023. 1

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 2

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 1

[5] Seyed Mohammad Hassan Erfani, Zhenyao Wu, Xinyi Wu, Song Wang, and Erfan Goharian. Atlantis: A benchmark for semantic segmentation of waterbody images. *Environmental Modelling & Software*, 149:105333, 2022. 1

[6] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776. IEEE, 2020. 1

[7] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your vit is secretly an image segmentation model. In *CVPR*, pages 25303–25313, 2025. 1

[8] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 1, 2

[9] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 1, 2

[10] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 1

[11] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *ICCV*, 2023. 1

[12] Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. *arXiv preprint arXiv:2411.09219*, 2024. 1, 2

[13] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 1

[14] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *ACM MM*, pages 506–515, 2021. 1