Unveiling the Knowledge of CLIP for Training-Free Open-Vocabulary Semantic Segmentation

Yajie Liu[1 2], **Guodong Wang**[1 2], **Jinjin Zhang**[1 2], **Qingjie Liu**[2], **Di Huang**[1 2 *]

[1] State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing 100191, China
[2] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{yajie_liu, wanggd, jinjin.zhang, qingjie.liu, dhuang}@buaa.edu.cn

Abstract

Training-free open-vocabulary semantic segmentation aims to explore the potential of frozen vision-language models (VLM) for segmentation tasks. Recent works reform the inference process of CLIP and utilize the features from the final layer to reconstruct dense representations for segmentation, demonstrating promising performance. However, the final layer tends to prioritize global components over local representations, leading to suboptimal robustness and effectiveness of existing methods. In this paper, we propose CLIPSeg, a novel training-free framework that fully exploits the diverse knowledge across layers in CLIP for dense predictions. Our study unveils two key discoveries: Firstly, the features in the middle layers exhibit high locality awareness and feature coherence compared to the final layer, based on which we propose the coherence enhanced residual attention module that generates semantic-aware attention. Secondly, despite not being directly aligned with the text, the deep layers capture valid local semantics that complement those in the final layer. Leveraging this insight, we introduce the deep semantic integration module to boost the patch semantics in the final block. Experiments conducted on 9 segmentation benchmarks with various CLIP models demonstrate that CLIPSeg consistently outperforms all training-free methods by substantial margins, *e.g.*, a 7.8% improvement in average mIoU for CLIP with a ViT-L backbone, and competes with learning-based counterparts in generalizing to novel concepts in an efficient way.

Introduction

Open-vocabulary semantic segmentation (OVSS) aims to segment the visual elements of arbitrary categories. Unlike conventional semantic segmentation (Zhou et al. 2019; Zhang et al. 2020) that only works on limited predefined categories, OVSS is more practical for real-world applications where infinite visual concepts exist. However, the high cost of dense annotations makes it impractical to scale segmentation to large sets of concepts in a fully supervised manner. Motivated by the strong generality and open-vocabulary understanding capabilities of CLIP (Radford et al. 2021), recent works on OVSS explore transferring the image-text alignment towards finer-granularity for pixel-level prediction.

*Corresponding author.

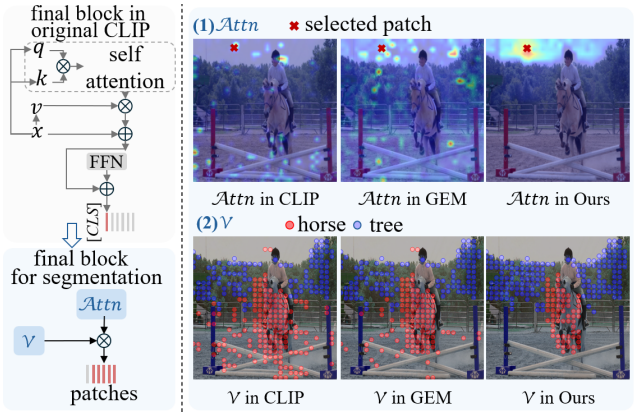

Figure 1: Comparison of our $\mathcal{A}ttn$ and $\mathcal{V}$ to previous works. For segmentation, the dense output of the final block in CLIP is constructed using the attention and the value. In current works (represented as GEM), the $\mathcal{A}ttn$ exhibits confusing relations among local patches and the patch semantics in $\mathcal{V}$ encounter contamination by the dominant category ("horse"). In contrast, our method generates semantic-aware attention $\mathcal{A}ttn$ and achieves accurate patch semantics in $\mathcal{V}$.

A group of literature integrates trainable modules into CLIP and relies on additional supervision to learn OVSS, such as dense annotations on limited categories (Xu et al. 2022) and additional image-text pairs (Cha, Mun, and Roh 2023). However, training with limited supervision inevitably compromises the generalization capability of the model, leading to underwhelming performance on unseen classes. Furthermore, the complex additional modules and the training process incur huge computational cost for OVSS.

In contrast, the training-free methods eliminate the need for fine-tuning while retain the capabilities and generalization of CLIP. CLIP-Surgery (Li et al. 2023) excludes the last feed forward (FFN) as it pushes the patch features towards negative semantics. ClearCLIP (Lan et al. 2024) removes the final residual connection to address the issue of noisy segmentation masks. With above modifications, the final block in CLIP for segmentation is reformed, as depicted in the left part of Fig. 1. In essence, the segmentation performance is primarily determined by the attention module (denoted as

$\mathcal{A}ttn$) and the value (denoted as $\mathcal{V}$).

On one hand, the original self-attention module in CLIP tends to focus on irrelevant semantic regions as illustrated in Fig. 1. In response to this issue, CLIP-Surgery proposes the use of value-value attention to enforce high attention on the patch itself. By integrating supplementary features (*e.g.*, query) from the final layer, GEM (Bousselham et al. 2024) further extends it to self-self attention (SSA) mechanism, which is widely adopted in subsequent works (Shao et al. 2024; Lan et al. 2024). However, as illustrated in the right part of Fig. 1, the resulting attention maps still exhibit confusing relations among local patches, erroneously assigning high weights to irrelevant regions, thereby resulting in the aggregation of unrelated semantics. On the other hand, the original final value embedding v_L in CLIP is commonly utilized as $\mathcal{V}$ in current works (Shao et al. 2024; Lan et al. 2024). Nevertheless, the patch semantics in v_L are partially contaminated by the dominated category (*e.g.*, "horse" in Fig. 1), which would lead to inaccurate category predictions for these patches and their relatives. In addition, CLIP-Surgery and GEM empirically employ multiple immediate value embeddings from deep layers as $\mathcal{V}$ without a clear analysis of the underlying mechanism, limiting further improvements.

To address these limitations, we introduce CLIPSeg, a novel training-free framework that explores the intrinsic knowledge at different layers in CLIP to construct $\mathcal{A}ttn$ and $\mathcal{V}$ for OVSS. We speculate that the confusing relationships in existing $\mathcal{A}ttn$ is attributed to the features used for construction, which typically come from the final layer. However, the final layer tends to prioritize global components, while disregard local representations. This motivates us to raise an intriguing question: ***Can leveraging features from shallow or middle layers result in higher quality attention?*** To construct an accurate $\mathcal{A}ttn$, the features need to achieve both intra-class compactness and inter-class separability. We quantify the intra-image feature coherence inspired from STEGO (Hamilton et al. 2022) and discover that **middle layers consistently exhibit better feature coherence compared to the final layer across various CLIP models, facilitating the generation of accurate attention maps.** Building upon this, we introduce the **C**oherence enhanced **R**esidual **A**ttention (CRA) module which identifies the optimal feature to construct the $\mathcal{A}ttn$.

In addition, we conjecture that the semantic contamination in current $\mathcal{V}$ (Lan et al. 2024) arises from the self-attention aggregation, which diffuses global information into local patches in the deep layers. To address this issue, we exclude the self-attention module during the forward process of deep layers and regenerate their dense outputs, denoted as $\widetilde{v}$. Subsequently, we re-use the value projection of the final layer to project $\widetilde{v}$ into the unified visual latent space to facilitate its semantic prediction. We conduct a quantitative analysis on $\widetilde{v}$ and discover that the patch semantics in $\widetilde{v}$ effectively complement the final value and alleviate the semantic contamination. Based on this, we propose the **D**eep **S**emantic **I**ntegration (DSI) module to boost the patch semantics of $\mathcal{V}$.

We apply the proposed framework to 8 widely-used CLIP models, namely CLIP, OpenCLIP, MetaCLIP, with both ViT-B (denoted as -B) and ViT-L (-L) backbones. The MetaCLIP is trained with different data scales, including 400M (denoted as -400M) and 5B image-text pairs (-5B). Extensive experiments conducted on 9 semantic segmentation benchmarks show that our method consistently achieves the state-of-the-art performance. Additionally, our training-free solution even outperforms some methods which are trained with additional image-text pairs by significant margins.

We highlight our contributions and discoveries as follows:

- We decouple the $\mathcal{A}ttn$ and $\mathcal{V}$ of the final block and reformulate them by exploring the characteristics of different layers.
- We propose the **C**oherence enhanced **R**esidual **A**ttention (CRA) module for $\mathcal{A}ttn$, inspired by the observation that features in the middle layer exhibit better coherence than those in the final layer of CLIP.
- We introduce the **D**eep **S**emantic **I**ntegration (DSI) module to construct $\mathcal{V}$, based on the observation that deep layers capture informative patch semantics that complement the final value, which is often contaminated by the dominant category.
- We comprehensively assess our method on 9 benchmarks, and the results clearly show that our approach significantly surpasses state-of-the-art methods.

Related Work

Open-vocabulary Semantic Segmentation. Motivated by the superior generality of large-scale vision-language models (Jia et al. 2021; Radford et al. 2021), prior works on OVSS (Ren et al. 2022; Liang et al. 2023; Xing et al. 2023; Zhang et al. 2023) explore methodologies to transfer the global cross-modal alignment capabilities towards finer granularity. Existing works can be broadly categorized into three groups based on the supervision paradigm. **Fully-supervised** methods (Xu et al. 2022, 2023) introduce dense annotations into training, typically using the COCO-Stuff dataset. However, training on limited categories leads to underwhelming performance on unseen classes and the reliance on dense annotations impose limitations on the scalability of the methods. To encompass a wide range of categories during training, **weakly-supervised** methods (Luo et al. 2023; Wang et al. 2024) rely on the readily available image-text paired data, such as the CC12M dataset. However, the training data scale is smaller compared to that used for CLIP, which compromises the generalization of the model. Moreover, the additional training incurs a large computational cost. **Training-free** methods (Lan et al. 2024; Zhao et al. 2024) explore the potential of frozen CLIP to generate segmentation masks. Our approach belongs to training-free open-vocabulary semantic segmentation.

Training-free Open-vocabulary Semantic Segmentation. We category current training-free methods into two groups based on their inference pipeline to generate segmentation masks. **Two-stage** methods (Sun et al. 2024; Luo et al. 2024) typically rely on Grad-CAM to generate mask proposals, which are then feed into CLIP for category predictions. However, the generated masks tend to capture only

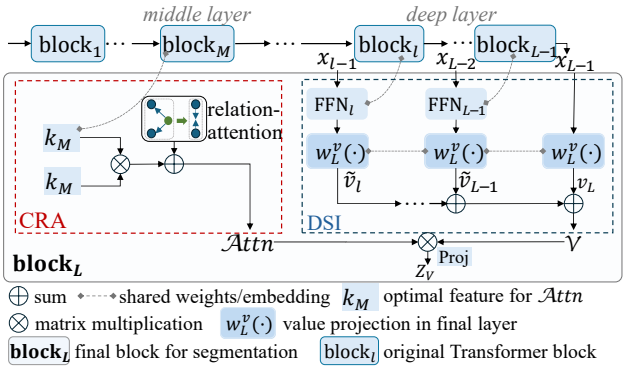

Figure 2: Overall pipeline and the final block of CLIPSeg. The CRA module identifies the optimal k_M, which is typically observed in the middle layer, for constructing $\mathcal{A}ttn$. The DSI module regenerates the dense outputs $\tilde{v}$ for deep layers and combines them with the final v_L to construct $\mathcal{V}$.

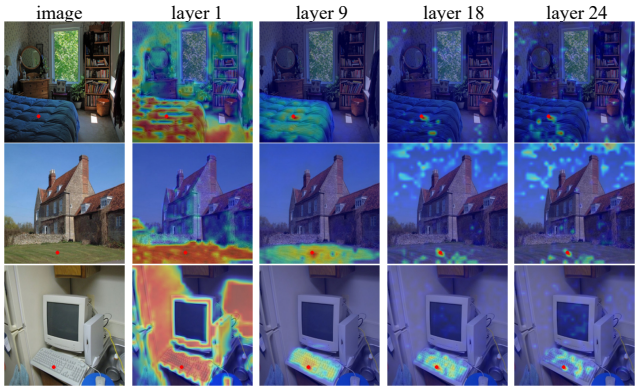

Figure 3: Comparison of patch response heatmaps generated by the k-k attention at different layers in CLIP-L. The red dot in the picture represents the selected patch. The final layer (24) exhibits confusing relationships with irrelevant regions, while the middle layer (9) leads to accurate attention on similar patches.

the most discriminative regions of the target, which necessitates iterative refinement to improve the quality of the masks. **One-stage** methods focus on reforming the inference stream of the visual encoder and conduct pixel-wise classification. MaskCLIP (Zhou, Loy, and Dai 2022) excludes the self-attention module in the last layer of CLIP and reveals that the final value embedding captures local representations that align well with text. CLIP-Surgery (Li et al. 2023) introduces a dual path which employs the value-value attention to replace the original self-attention that attends to opposite semantic regions. GEM (Bousselham et al. 2024) and SCLIP (Wang, Mei, and Yuille 2023) further extend the value-value attention to self-self attention (*e.g.*, query-query) mechanism. NACLIP (Hajimiri, Ayed, and Dolz 2024) and ClearCLIP (Lan et al. 2024) remove the final feed forward layer and residual connection since they lead to noisy segmentation masks.

Existing works on training-free OVSS primarily rely on features solely of the final layer to generate the dense output. However, the final layer tends to prioritize global components and fall short in providing the local knowledge required for semantic segmentation. Consequently, these methods may exhibit suboptimal robustness and effectiveness, especially when applied to CLIP models with a ViT-L backbone that contains a large number of layers. In this paper, we explore and harness the diverse knowledge captured across different layers in CLIP for dense predictions and propose a robust training-free framework for OVSS.

Method

Problem Definition

The visual encoder in CLIP with the ViT architecture comprises L Transformer blocks. And the final output can be described as follows:

$$\begin{aligned} x_L &= x_{L-1} + \text{Proj}(\text{softmax}(Attn) \cdot v_L), \\ x_L &= x_L + \text{FFN}(\text{LN}(x_L)), \end{aligned} \tag{1}$$

where x_{L-1}, v_L denote the output of the penultimate layer and the last value embedding, respectively. Proj denotes a projection layer and LN represents the layer normalization. In classification tasks, the $Attn$ is typically computed as the query-key (q-k) attention which aggregates the global information. To enhance the local representations required for segmentation, certain modifications have been introduced to Eq. 1. Following CLIP-Surgery (Li et al. 2023) and ClearCLIP (Lan et al. 2024), we exclude the last FFN which tends to push the patch features towards inverse semantics and ignore the final residual connection which leads to noisy segmentation masks. Consequently, the dense output $Z_v \in \mathbb{R}^{P \times D}$ of the visual encoder can be reformed as follows:

$$Z_v = \text{Proj}(\text{softmax}(\mathcal{A}ttn) \cdot \mathcal{V}), \tag{2}$$

where P, D represent the patch numbers and the dimension size, respectively. $[\cdot]$ stands for the dot production.

As illustrated in Eq. 2, the dense output Z_v is constructed by aggregating the patch semantics in $\mathcal{V}$ using $\mathcal{A}ttn$. Based on this, our analysis and enhancements for OVSS focus on two pivotal components: (1) We introduce the coherence enhanced residual attention (CRA) module for $\mathcal{A}ttn$, which robustly assigns high weights to similar patches while disregarding irrelevant ones. (2) We propose the deep semantic integration (DSI) module to boost the patch semantics captured by $\mathcal{V}$. The overall pipeline of the proposed method is presented in Fig. 2.

Coherence Enhanced Residual Attention

The $\mathcal{A}ttn$ determines the weights assigned to each token when constructing the feature of a patch. In existing works, features from the final layer are commonly utilized to compute self-self attention (*e.g.*, q-q, k-k) for $\mathcal{A}ttn$. However, as depicted in the last column of Fig. 3, the resulting attention maps present confusing relationships among patches, which may be attributed to the global information entailed in the patch features of the final layer. This motivates us to explore the shallow and middle layers for attention construction.

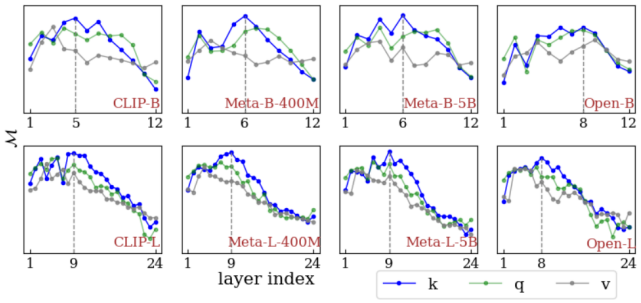

Figure 4: Comparison of $\mathcal{M}$ at different layers across 8 CLIP models. The dotted line in the graph indicates the layer index corresponding to the optimal feature for $\mathcal{A}ttn$. The features in the middle layers consistently demonstrate better coherence compared to the final layer. This trend is particularly pronounced in CLIP with a ViT-L backbone.

In order to construct accurate attention, it is crucial to ensure that the features exhibit both intra-class compactness and inter-class separability. This allows for the emphasis on semantically similar regions while disregarding unrelated ones. To this end, we develop the intra-image visual feature coherence metric, denoted as $\mathcal{M}$, to assess the features across layers for attention construction. For simplicity, we concatenate the output from multi-head attention to form a single query q, key k and value v for each layer. We take the key $k_i \in \mathbb{R}^{P\times D}$, where $i = 1, ..., L$ (the number of Transformer blocks), to illustrate the metric. Given the L2 normalized k_i, we start with computing the affinity matrix $S \in \mathbb{R}^{P\times P}$ among the patches within an image as $S = k_i \cdot k_i^T$. Subsequently, we employ the S as a binary classifier to predict whether two randomly selected patches p_1, p_2 belong to the same semantic category. The classification is conducted by applying a threshold to the similarity value between two patches. The ground truth for this binary classification task is derived from the semantic segmentation mask m, where the label is assigned as 1 if $m_{p_1} = m_{p_2}$ and 0 otherwise. The AUROC score of the prediction task is adopted as the metric $\mathcal{M}$. We employ the COCO-Stuff dataset to evaluate the intra-image feature coherence. Notably, we do not use the ground truth labels to tune any parameters of CLIP. By plotting the $\mathcal{M}$ of features across different layers in Fig. 4, we obtain the following observations:

- The feature coherence $\mathcal{M}$ in CLIP first increases and then decreases as the layers go deeper. The feature with the highest coherence, denoted as k_M, is typically observed in the middle layers, *e.g.*, layers 8-9 for CLIP with a ViT-L backbone.
- The features in the final layer demonstrate poor coherence, particularly noticeable in CLIP models with a ViT-L backbone that contains a large number of layers.

Building upon this, the CRA employs the optimal feature k_M for attention construction. The conventional approach is to exploit the self-self attention (SSA) mechanism. However, SSA solely focuses on the pairwise relationships between two patches, neglecting their relationships with other

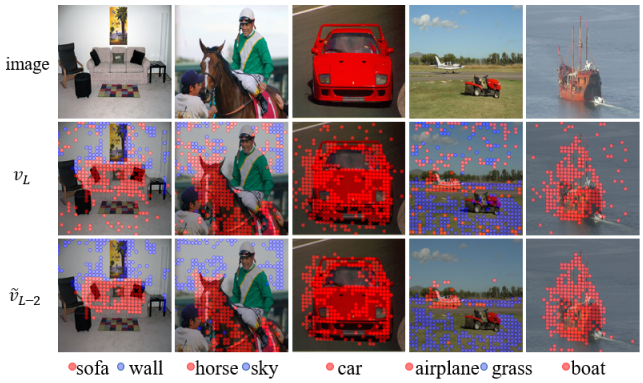

Figure 5: Comparison of the patch semantics in the final v_L and the $\widetilde{v}_{L-2}$ from the three-to-last layer in Meta-L-400M. The categories assigned to certain patches in the image are generated by matching their embedding to the class names in the Pascal Context-59 dataset. Not every patch is labeled with its category to provide a clear view. The patch tokens in the final v_L are partially contaminated by the dominated category which is indicated by the red dot in each image, while the $\widetilde{v}_{L-2}$ mitigates this phenomenon.

patches. In our CRA design, the similarity between two patches is not solely determined by their direct interactions, but is also influenced by their respective relationships with a third-party one, which can be formulated as follows:

$$\mathcal{A}ttn = (k_M \cdot k_M^T + \mathcal{G}(k_M))\tau, \tag{3}$$

where τ is the temperature hyperparameter. $\mathcal{G}$ stands for the relation-attention module, which is accomplished by establishing connections between two points with weights ϵ if both of their similarities to common third patches exceed the threshold value ϵ. In practice, $\mathcal{G}$ can be effectively achieved using the soft-clustering methods.

Deep Semantic Integration

The $\mathcal{V}$ provides the foundational patch semantics for constructing the dense representations. In previous works (Shao et al. 2024; Lan et al. 2024), the final value v_L is typically utilized as $\mathcal{V}$. However, as depicted in the second row of Fig. 5, the patch semantics in v_L are partially contaminated by the dominant category, leading to inaccurate category predictions for these patches and their relatives. We conjecture such phenomenon is originated from the q-k attention aggregation, which diffuses the global information into patch tokens in the deep layers. To verify this hypothesis, we exclude the self-attention and regenerate the dense outputs $\widetilde{v}_i$ for the previous layers. The process can be defined as follows:

$$\begin{aligned} \widetilde{x}_i &= x_{i-1} + \text{FFN}(x_{i-1}) \\ \widetilde{v}_i &= w_L^v \cdot \widetilde{x}_i \end{aligned} \quad i = 1, ..., L-1, \tag{4}$$

where FFN denotes the feed forward in i-th layer. Notably, the value projection w_L^v in the final layer is employed to project the intermediate dense output into the final visual latent space to facilitate the semantic prediction. As shown

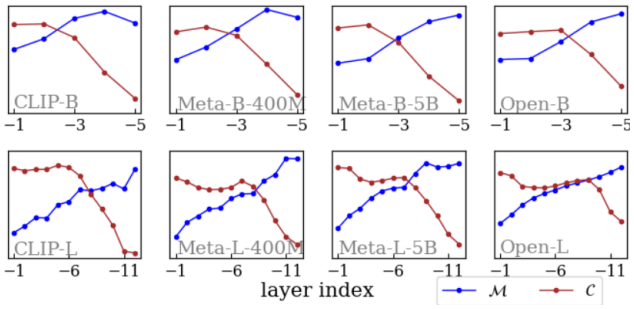

Figure 6: Comparison of the local representations captured by v_L and $\widetilde{v}_i$. The v_L corresponds to the layer index -1, representing the final layer. The layer index -2 refers to the $\widetilde{v}_i$ generated by the penultimate layer. As the layer becomes shallower, the feature coherence $\mathcal{M}$ increases while the cross-modal semantic alignment $\mathcal{C}$ decreases.

in Fig. 5, $\widetilde{v}_i$ mitigates the semantic contamination of patch tokens by the dominant category.

However, the $\widetilde{v}_i$ derived from previous layers may suffer from inaccurate semantic alignment due to the information loss during the reformed forward process. To evaluate the quality of the local representations captured by $\widetilde{v}_i$ and v_L, we assess them from two perspectives: (1) locality awareness, which measures whether the patch token accurately captures its local information. The occurrence of patches contaminated by global semantics would lead to confusing relationships and poor locality awareness. Therefore, the coherence $\mathcal{M}$ is employed here. (2) Cross-modal Semantic Alignment $\mathcal{C}$, which reflects the alignment between local visual representations and text embeddings. The average pixel accuracy on COCO-Stuff dataset is employed as $\mathcal{C}$. Low semantic alignment may lead to inaccurate category predictions. Fig. 6 illustrates the evaluation results across 8 CLIP models, from which we can make the following observations:

- The final value v_L exhibits poor locality awareness but strong cross-modal alignment.
- As the layer becomes shallower, the locality awareness of $\widetilde{v}_i$ increases while the cross-modal semantic alignment $\mathcal{C}$ decreases. Notably, in most models, there is a significant drop in $\mathcal{C}$ at the fourth-to-last layer.

In order to achieve a trade-off between locality awareness and cross-modal semantic alignment within the local representations, we introduce the deep semantic integration (DSI) module which combines the $\widetilde{v}_i$ and v_L to generate $\mathcal{V}$, as illustrated in the Fig. 2. We denote the layer index of the start layer to be combined as l. It is worth nothing that our DSI module differs from the dual-path in CLIP-Surgery and GEM which generates the dense output for previous layers via $\widetilde{v}_i = w_i^v \cdot x_i$. They directly employ the immediate value embeddings in deep layers that would encounter more information loss and a greater decrease in the semantic alignment. Additional analysis are provided in the appendix.

Experiments

Implementation Details

We apply the proposed framework on 8 CLIP models for OVSS, including CLIP-B/L (Radford et al. 2021) released by OpenAI, and Open-B/L (Cherti et al. 2023) pretrained on LAION-400M (Schuhmann et al. 2021), and Meta-B/L-400M/5B (Xu et al. 2024). Among the eight models, the optimal features k_M for $Attn$ consistently correspond to the key embedding from a certain layer. The layer index is determined according to Fig. 4. For simplicity and without loss of generality, we employ the soft-dbscan approach to calculate the relation-attention. Specifically, if two points are density-reachable within a cosine distance threshold ϵ, their relation-attention is assigned ϵ, otherwise 0. The cosine distance threshold ϵ is empirically set to 0.75. The temperature in Eq. 3 is set to 6 across all models and datasets. If not specified, we ensemble the dense output of last three layers to construct $\mathcal{V}$.

Comparison to State-of-the-art Methods

Evaluation protocol. We follow the widely-used evaluation protocol, as introduced in TCL (Cha, Mun, and Roh 2023) to evaluate our method across 9 segmentation benchmarks in a zero-shot manner. These benchmarks are categorized into two groups: (i) without background class including Pascal VOC20 (Everingham et al. 2010) with 20 classes (denoted as V20), Pascal Context (Mottaghi et al. 2014) with 459 classes in the full version (C459) and the most frequent 59 classes in the C59 version, COCO-Stuff (Caesar, Uijlings, and Ferrari 2018) with 171 classes (STUFF), ADE20k (Zhou et al. 2019) with 847 classes in the full version (A847) and A150 version with the most frequent 150 classes and Cityscapes (CITY) (Cordts et al. 2016) with 19 classes. (ii) with a background class including Pascal Context 60 (C60) and COCO object with 80 classes (COCO).

Comparison to training-free methods. We compare our method with state-of-the-art training-free approaches including MaskCLIP (Zhou, Loy, and Dai 2022), SCLIP (Wang, Mei, and Yuille 2023), GEM (Bousselham et al. 2024), CLIPtrase (Shao et al. 2024) and ClearCLIP (Lan et al. 2024). In cases where results for various models have not been reported by the authors, we generate these results using their official released code. Notably, when evaluating on benchmarks without background class, we exclude class name-based tricks that introduce additional synonyms in SCLIP to ensure a fair comparison. Following previous works (Wang, Mei, and Yuille 2023), we incorporate pixel-adaptive mask refinement (PAMR) (Araslanov and Roth 2020) into all methods for a fair comparison.

As demonstrated in Tab. 1, we consistently achieve state-of-the-art performance on all benchmarks across various CLIP models, surpassing previous methods by significant margins. **Notably**, existing methods obtain lower performance for CLIP models with a ViT-L backbone compared to the ViT-B counterpart. The counter-intuitive phenomena can be attributed to the significantly inferior feature co-

Methods	VLM	C60	COCO	V20	CITY	STUFF	A150	A847	C59	C459	Avg.
MaskCLIP [ECCV22]	CLIP-B	22.6	18.9	72.1	11.2	15.1	9.0	-	25.3	-	-
SCLIP [arxiv23]		31.5	31.2	79.4	34.1	23.9	17.8	5.3	36.1	9.6	29.9
GEM [CVPR24]		33	33.7	73	30.8	24.2	15.6	6.2	35.1	11.2	29.2
CLIPtrase [ECCV24]		31.1	33.4	80.6	18.6	23.9	17.5	5.6	35.7	10.1	28.5
ClearCLIP [ECCV24]		32.8	33.3	74.1	29.7	24.5	16.9	6.2	35.2	10.6	27.2
CLIPSeg (Ours)	CLIP-B	**35.3**	**37.9**	**83.5**	**36.4**	**27.1**	**19.7**	**6.9**	**39.7**	**11.5**	**33.1** (+3.2)
SCLIP [arxiv23]	CLIP-L	25.1	27.2	74	23.8	18.2	12.1	3.8	27.4	8.1	24.4
GEM [CVPR24]		25.6	32.3	75.5	26.4	19.4	12.8	5.4	27.6	9.5	26.0
CLIPtrase [ECCV24]		24.4	27.9	65.2	17.4	18.1	12.4	5.1	27.4	8.5	22.9
ClearCLIP [ECCV24]		27.3	32.3	74.1	29.1	20.7	15.3	6.3	29.5	10.4	27.9
CLIPSeg (Ours)	CLIP-L	**37.3**	**40.3**	**85.6**	**41.2**	**28.7**	**23**	**9.1**	**42.4**	**13.6**	**35.7** (+7.8)
SCLIP [arxiv23]	Meta-L-5B	25.5	26.1	71.7	21.7	17.3	13.2	4.6	28.2	8.6	24.1
GEM [CVPR24]		28.9	33.8	77	27.9	22.5	16.1	6.4	32.5	11.2	28.5
CLIPtrase [ECCV24]		27.5	31.1	71.5	14.7	20.9	15.8	6.4	30.8	10.8	25.5
ClearCLIP [ECCV24]		28.3	31.1	75	26.9	22.1	17.3	6.9	32.2	11.4	27.9
CLIPSeg (Ours)	Meta-L-5B	**39.4**	**41.9**	**87.2**	**41.8**	**30.3**	**25.9**	**9.9**	**45.1**	**15**	**37.4** (+8.9)

Table 1: Zero-shot segmentation performance comparison to training-free methods. The generated masks for all methods are further refined using the post-processing method PAMR. The metric mIoU (%) is used in every experiment. We highlight the **best** and second-best results.

Methods	V20	A150	CITY	C59	STUFF	Avg.
MaskCLIP [ECCV22]	74.9	9.8	22.3	26.4	16.4	30
SCLIP [arxiv23]	77.9	16.5	32.3	34.5	22.8	36.8
GEM [CVPR24]	79.9	15.7	31.4	35.9	23.7	37.3
ClearCLIP [ECCV24]	80.9	16.7	30.4	35.9	23.9	37.6
CLIPtrase [ECCV24]	81.2	17	18.4	34.9	24.1	35.1
CLIPSeg (Ours)	**83.2**	**19.2**	**35.9**	**38.7**	**26.4**	**40.7**

Table 2: Zero-shot segmentation performance comparison to training-free methods for CLIP-B without post-processing.

herence observed in the final layer of ViT-L based CLIP models (as shown in Fig. 4). Our approach addresses this limitation by capturing diverse knowledge across layers, outperforming the high-performing ClearCLIP by **7.8**% in average mIoU for CLIP-L and achieving an impressive **8.9**% mIoU improvement compared to GEM for Meta-L-5B. We additionally provide a performance comparison without post-processing in Tab. 2, where our method consistently achieves state-of-the-art performance. The qualitative comparison of our method with state-of-the-art approaches is presented in Fig. 7.

Comparison to training-based methods. To further demonstrate the effectiveness of CLIPSeg, we conduct a comparison with methods that rely on intricate training process on datasets with numerous novel concepts, including A847 and C459. These approaches can be categorized into two groups based on the supervision paradigm: (1) **Fully-supervised** methods rely on dense annotations, typically training on the COCO-Stuff dataset with 156 classes. These methods often require the use of additional models specifically designed for segmentation, such as mask proposal generation. (2) **Weakly-supervised** methods train on additional image-text pairs for fine-grained alignment.

Methods	Training Datadet	Extra Module	A847	C459
Zegformer [CVPR22]	STUFF-156	✓	4.9	9.1
OVSeg [CVPR23]	STUFF-156	✓	7.0	10.4
DeOP [ICCV23]	STUFF-156	✓	**7.1**	9.4
SegCLIP [ICML23]	CC3M+CC	✓	3.0	5.7
TCL [CVPR23]	CC3M&12M	✓	6.2	8.9
CoDe [CVPR24]	CC3M&12M	✓	6.1	10.0
CLIPSeg (Ours)	✗	✗	6.9	**11.5**

Table 3: Zero-shot segmentation performance comparison to training-based methods using CLIP-B. CC denotes the COCO Caption dataset.

As demonstrated in Tab. 3, our approach consistently outperforms the weakly-supervised methods that rely on addition training by significant margins. More importantly, our method achieves comparable results on A847 and outperforms fully-supervised counterparts on C459. For instance, we achieve 11.5% mIoU on C459 using CLIP-B in a training-free manner while DeOP attains 9.4 % mIoU even with additional training on COCO-Stuff-156.

Ablation Study

In this section, we conduct ablation studies to evaluate the effects of core components of the proposed method. **Notably**, all experiments are conducted based on Eq. 2, following the modifications that exclude the last FFN and residual connection. The post-processing methods are not employed.

Effect of the coherence enhanced residual attention. To demonstrate the effectiveness of our CRA module to the overall segmentation performance, we present the results

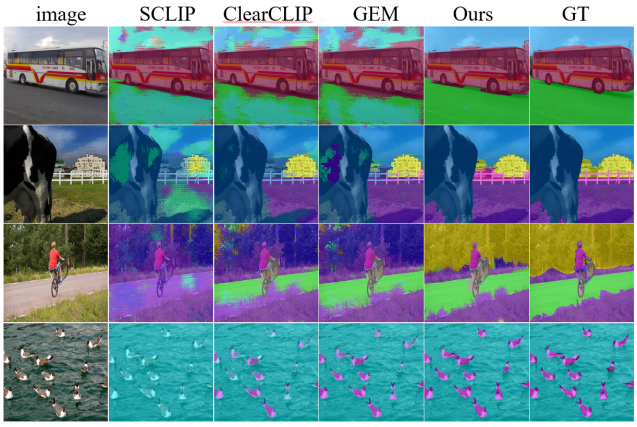

Figure 7: Qualitative comparison results on C59 for Meta-L-5B. The official palette from MMSegmentation is utilized.

$\mathcal{A}ttn$	VLM	A150	CITY	C59	STUFF	Avg.
q-k	CLIP -B	8.9	11.1	20.4	14	13.6
k_L-k_L		15.8	21.4	35.4	24.2	24.2
k_M-k_M (Ours)		18.9	35.1	38.4	26.1	29.6
+ $\mathcal{G}(k_M)$ (Ours)		19.2	35.9	38.7	26.4	**30.1**
q-k	CLIP -L	10.6	17.1	23.4	15.6	16.7
k_L-k_L		17.3	28.7	34.9	23.7	26.2
k_M-k_M (Ours)		22.1	39.8	41	27.6	32.6
+ $\mathcal{G}(k_M)$ (Ours)		22.3	40.1	41.3	28	**32.9**

Table 4: Ablation studies on the $\mathcal{A}ttn$. The k_L denotes the key embedding in the last layer. And k_M represents optimal feature for $\mathcal{A}ttn$ from a certain middle layer. "+ $\mathcal{G}(k_M)$" extends the k_M-k_M with relation-attention, which corresponds to the complete CRA module.

with different $\mathcal{A}ttn$ in Tab 4. The baseline refers to the original self-attention module that performs q-k attention. The line k_L-k_L represents the use of k in the final layer, as previously done in works (Bousselham et al. 2024; Shao et al. 2024), with the SSA mechanism to construct $\mathcal{A}ttn$. The line k_M-k_M indicates that the optimal embedding from a certain middle layer is utilized to construct the SSA. The line with "+ $\mathcal{G}(k_M)$" extends k_M-k_M by incorporating relation-attention, which corresponds to the complete CRA module.

As illustrated in Tab. 4, when employing the SSA mechanism to construct $\mathcal{A}ttn$ for CLIP-B, the k_M achieves an improvement of 5.4% in average mIoU compared to the k_L. For CLIP-L, the k_M outperforms k_L by a significant margin with 6.4% mIoU improvement, which could be attributed to the poor feature coherence of k_L in CLIP-L . Moreover, the relation-attention, which enhances the SSA by incorporating patch relationships, leads to an additional improvement of 0.5% mIoU for CLIP-B.

Effect of the DSI. Tab. 5 presents the ablation studies on the DSI. The l represents the layer index of the start layer to be combined in the DSI module. The "-1" baseline indicates the sole utilization of the final v_L embedding. As shown in Tab. 5, merging the local features for the last three

l	VLM	V20	A150	CITY	C59	STUFF	Avg.
-1	CLIP -B	82.5	18.8	34.1	37.9	25.7	39.8
-2		83.8	18.9	34.8	37.9	25.8	40.2
-3		83.2	19.2	35.9	38.7	26.4	**40.7**
-4		81.7	18.9	37.2	39	26.2	40.6
-1	META-L -400M	83.7	22.6	36.4	40	26.3	41.8
-2		85.2	23.2	39	41	27.5	43.2
-3		85.1	23.2	38.2	41.8	28	**43.3**
-4		84.3	22.9	37.7	41.8	27.7	42.9

Table 5: Ablation studies on the $\mathcal{V}$. The "-1" baseline indicates the sole utilization of the v_L embedding in the last layer. The l represents the layer index of the start layer to be combined in the DSI module. Our default settings in the main experiments are **bold**.

τ	VLM	V20	A150	CITY	C59	STUFF	Avg.
5	CLIP -B	83.6	18.5	31.7	38.1	26.1	39.6
6		83.2	19.2	35.9	38.7	26.4	**40.7**
7		82.2	19.3	37.5	38.6	26.1	40.7
8		80.7	19.1	37.9	38.2	25.5	40.3
5	CLIP -L	85.3	21.3	38.8	40.5	27.7	42.7
6		85.1	22.3	40.1	41.3	28	**43.4**
7		84.4	22.7	40.3	41.5	27.8	43.3
8		83.2	22.6	40.1	41.2	27.3	42.9

Table 6: Ablation studies on the temperature τ. Our default settings in the main experiments are **bold**.

layers consistently results in the best performance. It leads to a 0.9% average mIoU improvement for CLIP-B and a 1.5% mIoU improvement for META-L-400M. These results align with the observation in Fig. 6 that most models encounter a rapid drop in the semantic alignment at the four-to-last layer.

Effect of the temperature τ. We perform an ablation study on the performance impact of different values of τ in Tab. 6. The optimal values of τ for achieving the best performance vary across different datasets. Consistently, setting the value of τ to 6 yields the best performance in terms of average mIoU across various CLIP models.

Conclusion

In this paper, we propose CLIPSeg, a robust training-free open-vocabulary semantic segmentation framework with frozen CLIP models. We decouple the construction of dense output of CLIP into attention and patch semantics and emphasize the importance of exploiting features from different layers for their construction. We introduce the coherence enhanced residual attention module which generates the semantic-aware attention, leveraging the high coherence observed in features from middle layers. In addition, we propose the deep semantic integration module that enhances the local representations and mitigates the semantic contamination in the final value by incorporating dense outputs regenerated from deep layers. Extensive experiments conducted on 9 segmentation benchmarks demonstrate the superiority of our method.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (82441024, 62302031, 62176017), the Zhejiang Provincial Natural Science Foundation of China (LQ23F020024), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01020), the Research Program of State Key Laboratory of Critical Software Environment, and the Fundamental Research Funds for the Central Universities.

References

Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4253–4262.

Bousselham, W.; Petersen, F.; Ferrari, V.; and Kuehne, H. 2024. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3828–3837.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 1209–1218.

Cha, J.; Mun, J.; and Roh, B. 2023. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, 11165–11174.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 3213–3223.

Everingham, M.; van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Hajimiri, S.; Ayed, I. B.; and Dolz, J. 2024. Pay Attention to Your Neighbours: Training-Free Open-Vocabulary Semantic Segmentation. *arXiv preprint arXiv:2404.08181*.

Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; and Freeman, W. T. 2022. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. In *International Conference on Learning Representations*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 4904–4916. PMLR.

Lan, M.; Chen, C.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2024. ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference. In *ECCV*.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 7061–7070.

Luo, H.; Bao, J.; Wu, Y.; He, X.; and Li, T. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, 23033–23044. PMLR.

Luo, J.; Khandelwal, S.; Sigal, L.; and Li, B. 2024. Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4029–4040.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 891–898.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, P.; Li, C.; Xu, H.; Zhu, Y.; Wang, G.; Liu, J.; Chang, X.; and Liang, X. 2022. ViewCo: Discovering Text-Supervised Segmentation Masks via Multi-View Semantic Consistency. In *The Eleventh International Conference on Learning Representations*.

Schuhmann, C.; Kaczmarczyk, R.; Komatsuzaki, A.; Katta, A.; Vencu, R.; Beaumont, R.; Jitsev, J.; Coombes, T.; and Mullis, C. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*, FZJ-2022-00923. Jülich Supercomputing Center.

Shao, T.; Tian, Z.; Zhao, H.; and Su, J. 2024. Explore the Potential of CLIP for Training-Free Open Vocabulary Semantic Segmentation. *arXiv e-prints*, arXiv–2407.

Sun, S.; Li, R.; Torr, P.; Gu, X.; and Li, S. 2024. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13171–13182.

Wang, F.; Mei, J.; and Yuille, A. 2023. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*.

Wang, H.; Vasu, P. K. A.; Faghri, F.; Vemulapalli, R.; Farajtabar, M.; Mehta, S.; Rastegari, M.; Tuzel, O.; and Pouransari, H. 2024. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3635–3647.

Xing, Y.; Kang, J.; Xiao, A.; Nie, J.; Shao, L.; and Lu, S. 2023. Rewrite Caption Semantics: Bridging Semantic Gaps for Language-Supervised Semantic Segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xu, H.; Xie, S.; Tan, X.; Huang, P.-Y.; Howes, R.; Sharma, V.; Li, S.-W.; Ghosh, G.; Zettlemoyer, L.; and Feichtenhofer,

C. 2024. Demystifying CLIP Data. In *The Twelfth International Conference on Learning Representations*.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, F.; Zhou, T.; Li, B.; He, H.; Ma, C.; Zhang, T.; Yao, J.; Zhang, Y.; and Wang, Y. 2023. Uncovering Prototypical Knowledge for Weakly Open-Vocabulary Semantic Segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; and Chan, A. B. 2024. Gradient-based visual explanation for transformer-based clip. In *International Conference on Machine Learning*, 61072–61091. PMLR.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3): 302–321.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*, 696–712. Springer.