
Spatial Discriminability of CLIP for Training-Free Open-Vocabulary Semantic Segmentation

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Extending CLIP models to semantic segmentation remains a considerable chal-
2 lenge, largely due to the misalignment between their image-level pre-training
3 objectives and the pixel-level spatial understanding required for dense predictions.
4 Prior efforts have achieved encouraging results by reorganizing the final layer and
5 feature representations of CLIP to enhance dense predictions. However, these
6 approaches often inherit the global alignment bias of the final layer, leading to
7 suboptimal spatial discriminability and segmentation performance. In this work,
8 we propose TLH-CLIP, a novel training-free framework that systematically exploits
9 the spatial discriminability across *Token*, *Layer* and *Head* levels in CLIP for dense
10 predictions. Through comprehensive analysis, we uncover three key findings: (i)
11 some anomalous tokens emerges in the final layers, which are category-agnostic
12 but disproportionately attract attention from semantically meaningful patch tokens,
13 thereby degrading spatial discriminability; (ii) the final few layers primarily en-
14 hance global image-text alignment with great sacrifice of local discriminability
15 (e.g., last 3 layers in ViT-B-16 and 5 layers in ViT-L-14); (iii) a few attention heads
16 (e.g., 10 out of 144 in ViT-B/16) demonstrate strong spatial discriminability across
17 different datasets. Motivated by these insights, we propose three complementary
18 techniques: abnormal token replacement, semantic-spatial reweighting, and se-
19 lective head enhancement to effectively recover spatial coherence and improve
20 segmentation performance without any additional training, auxiliary pre-trained
21 networks, or extensive hyperparameter tuning. Extensive experiments on 8 com-
22 mon semantic segmentation benchmarks demonstrate that TLH-CLIP achieves
23 state-of-the-art performance across diverse scenarios, highlighting its effectiveness
24 and practicality for real-world deployment.

1 Introduction

26 Recent advances in vision-language pretrained models, such as CLIP [1], have demonstrated remark-
27 able generalization and open-vocabulary recognition capabilities at the image level, thereby opening
28 up possibilities for transferring image-text alignment to pixel-level tasks. Despite this progress, they
29 often underperform in dense prediction tasks like semantic segmentation, primarily due to their
30 limited capacity to localize fine-grained visual details [2, 3]. To address these limitations, several
31 studies have incorporated trainable modules into CLIP, typically relying on additional forms of
32 supervision such as dense annotations for a restricted set of categories [4, 5, 6, 7] or supplementary
33 image-text pairs [8, 9, 10, 11, 12]. Although these approaches have demonstrated improved seg-
34 mentation performance, they incur significant computational and annotation costs. Furthermore, the
35 dependence on limited supervision often undermines the generalizability of the model, making it
36 prone to overfitting the training distribution.

These challenges have sparked increasing interest in training-free methods[3, 13, 14, 15, 16, 17, 18, 19, 20], which aim to adapt CLIP’s pre-trained representations for semantic segmentation without additional training, while preserving its generalization capability. A key difficulty in this direction is enhancing spatial representations for accurate pixel-level predictions. For instance, MaskCLIP[14] computes similarity between key features in the final attention layer to enrich patch embeddings. SCLIP [3] replaces the standard query-key attention with correlative self-attention (query-query and key-key). ClearCLIP [15] further removes residual connections and discards the FFN in the final layer to reduce noise and improve spatial alignment. ResCLIP [20] incorporates attention maps from earlier layers to refine final-layer attention map. However, these methods largely focus on modifying the final-layer attention, often leading to suboptimal ambiguous local relationships and noisy segmentation. To address spatial limitations, some approaches incorporate features from auxiliary backbones such as DINO [21, 17], SAM [17, 22], or diffusion models [23, 24]. While effective, these methods incur significant computational and memory overhead.

Motivated by these limitations, we begin with a layer-wise analysis of spatial discriminability and text-semantic alignment within the CLIP model. As shown in Figure 1, we observe a clear spatial-semantic trade-off in the final layers: spatial discriminability drops sharply, while the improvement in semantic alignment is relatively marginal. To understand the cause of this phenomenon, we further examine internal token interactions and structural patterns across layers. Through attention map visualizations, we find that certain abnormal tokens emerge in the deeper layers, attracting disproportionately high attention from nearly all spatial positions. This behavior causes the majority of tokens to converge on a small subset, thereby disrupting the spatial coherence of the representation. Further analysis reveals that these abnormal tokens exhibit sparse and high-magnitude activations. Moreover, they are class-agnostic, as their similarity remains consistent across different positions, layers, and input samples, indicating a lack of semantic specificity. Contrary to prior assumptions that such tokens encode global semantic content, our findings suggest they may instead function as bias components that offset global-mean features, thereby facilitating alignment with text embeddings.

Based on the analysis, we propose TLH-CLIP, a training-free framework that leverages the inherent properties of CLIP to enhance the spatial discriminability of visual features while preserving their semantic alignment. TLH-CLIP comprises three complementary strategies: abnormal token replacement (ATR), spatial-semantic reweighting (SSR), and selective head enhancement (SHE). Specifically, the ATR employs hoyer scores to identify abnormal tokens by thresholding their characteristic sparsity. Once detected, these anomalous tokens are replaced with a weighted average of normal tokens, based on spatial distance. To mitigate the degradation of spatial discriminability in the earlier final layers, SSR reweights the contributions of the residual pathway relative to the attention and FFN submodules. This adjustment restores a better balance between spatial coherence and semantic abstraction, leveraging the fact that late-intermediate layers exhibit stronger spatial discriminability while maintaining comparable levels of semantic alignment. Finally, SHE further enhances spatial coherence by selectively aggregating features from attention heads with high spatial discriminability, using them to refine the output representations. Experimental results demonstrate that TLH-CLIP achieves significant performance improvements when integrated into various baseline methods, establishing new state-of-the-art results across eight benchmark datasets.

Contributions. Our contributions can be summarized as follows:

- We conduct a comprehensive analysis of spatial discriminability at the token, head, and layer levels.
- We propose, a novel training-free approach, terms TLH-CLIP. To the best of our knowledge, this is the first work to explicitly modify the inference procedure prior to the final layer, enabling improved spatial coherence without compromising semantic alignment.
- The extensive experiment results on open-vocabulary semantic segmentation tasks consistently demonstrate the effectiveness of the proposed method.

2 Analysis

2.1 Preliminaries

CLIP employs a Vision Transformer (ViT) [25] as its image encoder to generate visual representations that are aligned with corresponding textual descriptions. The vision encoder first tokenizes an input image of size $H \times W \times 3$ by dividing it into a grid of non-overlapping patches of size $P \times P$,

yielding $h = H/P$ rows and $w = W/P$ columns of patches. Each patch is then linearly projected into a D -dimensional embedding space, $\mathbf{x}_i \in \mathbb{R}^D$, and augmented with positional embeddings. An additional learnable [CLS] token is prepended to the sequence and is later used for image-level prediction. The resulting token sequence is denoted as $\mathbf{X}^0 = [\mathbf{x}_{\text{cls}}^0, \mathbf{x}_1^0, \dots, \mathbf{x}_{hw}^0] \in \mathbb{R}^{(1+hw) \times D}$. This sequence is passed through a stack of L Transformer encoder layers, each consisting of a multi-head self-attention (MSA) module followed by a feed-forward network (FFN). Let LN denotes layer normalization, the token representations are updated at each layer l as follow:

$$\hat{\mathbf{X}}^l = \mathbf{X}^{l-1} + \text{Attn}(\text{LN}(\mathbf{X}^{l-1})), \quad (1)$$

$$\mathbf{X}^l = \hat{\mathbf{X}}^l + \text{FFN}(\text{LN}(\hat{\mathbf{X}}^l)). \quad (2)$$

The CLIP model is originally trained on large-scale image-text pairs for open-vocabulary image recognition tasks. To extend it to semantic segmentation, a natural approach is to compute the similarity between the visual tokens $\mathbf{X}^L = [\mathbf{x}_1^L, \dots, \mathbf{x}_{hw}^L]$ from the final Transformer layer and the textual embeddings of C category names, denoted by $\mathbf{t} \in \mathbb{R}^{C \times D}$. This results in a patch-text similarity map of size $hw \times C$. Denote \mathbf{t}_c as the embedding of the c -th class name, the final segmentation prediction is obtained by applying an argmax operation over the class dimension of this similarity map, as follows:

$$\hat{c}(\mathbf{x}_i) = \arg \max_c \frac{\langle \mathbf{x}_i^L, \mathbf{t}_c \rangle}{\|\mathbf{x}_i^L\| \cdot \|\mathbf{t}_c\|}, \quad (3)$$

Ideally, for effective semantic segmentation, the vision encoder should produce feature representations that satisfy two key properties:

- **Spatial discriminability (SD):** token features should exhibit high internal consistency within the same semantic category while remaining clearly distinguishable from those of other categories, thereby enabling accurate and clean segmentation results.
- **Semantic alignment (SA):** token features should be well-aligned with their corresponding textual embeddings to enable semantically meaningful segmentation results.

Beyond their importance in open-vocabulary semantic segmentation, these two properties are also more highly relevant to the development of multimodal large language models (MLLMs), as the vision encoder of CLIP is often directly employed to extract visual representations without additional training, serving as input to downstream language models such as LLaVA [26, 27]. In this work, we aim to enhance the spatial discriminability of CLIP features in a training-free manner, thereby preserving the its strong generalization capability.

2.2 Analysis of layer-wise spatial discriminability and semantic alignment

Significant decline in SD with marginal gains in SA in the final layers. To assess whether CLIP visual features exhibit the desired properties, we investigate the layer-wise SD and SA within CLIP models. To quantitatively assess SD property, we follow the evaluation protocol proposed in [28]. In particular, we extract patch-level feature representations from the vision encoder for each image and associate them with corresponding semantic labels using the ground-truth segmentation masks from Pascal VOC [29], PASCAL Context [30], ADE20K [31], and COCO-Stuff [32] datasets. Specifically, let $\mathbf{x}_i^l \in \mathbb{R}^D$ and $\mathbf{x}_j^l \in \mathbb{R}^D$ denote the feature representations of two image patches i and j extracted from the l -th layer of the encoder. These feature vectors are ℓ_2 -normalized, and their cosine similarity is computed to serve as the prediction of a binary classifier that indicates whether the two patches belong to the same semantic category. Given the corresponding semantic labels $t(\mathbf{x}_i)$ and $t(\mathbf{x}_j)$, the target value for classification is set to 1 if $t(\mathbf{x}_i) = t(\mathbf{x}_j)$, and 0 otherwise. To evaluate the SA property, we extract the intermediate representations $\mathbf{x}_i^l \in \mathbb{R}^D$ from each individual visual token at layer l , and use them as inputs to the final layer to project these features into the final visual latent space for semantic prediction. Following [15], we remove the FFN and residual connections in the final layer to avoid introducing contaminating semantic information. Additionally, inspired by [14], we replace the last-layer attention matrix with an identity matrix to avoid noisy integration during the final attention computation. The final visual representation of each layers can be expressed as $\mathbf{v}_i^l = \mathbf{x}_i^l \mathbf{W}_v^L \mathbf{W}_o^L \in \mathbb{R}^D$, where \mathbf{W}_v^L and \mathbf{W}_o^L denotes the value and output project matrix in last-layer MSA module. Based on these representations, SA is measured using the average accuracy between the predicted and ground-truth semantic labels, following Equation (3).

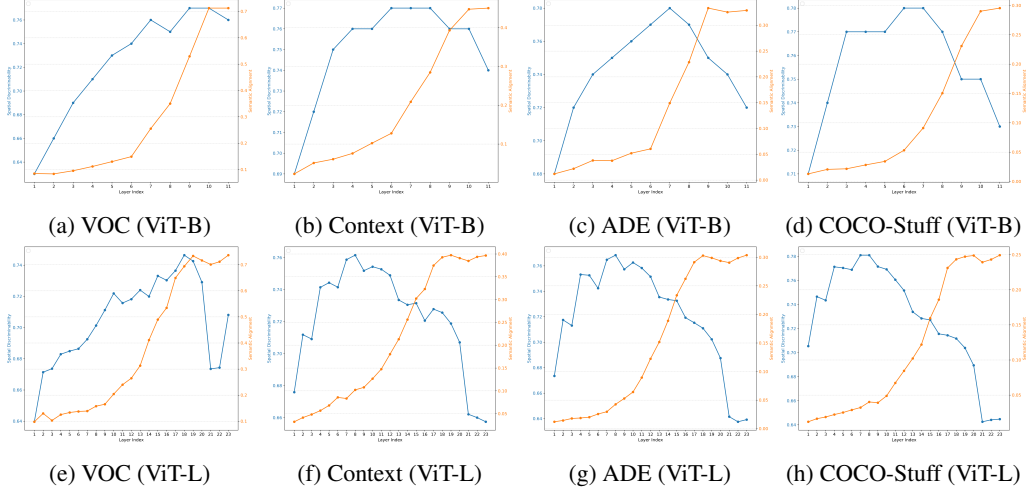


Figure 1: Layer-wise analysis of spatial discriminability (blue curves) and semantic alignment (orange curves) within the CLIP vision encoders across different datasets. The final layer is excluded from the analysis to avoid discrepancies caused by prior modifications to the last-layer in different methods.

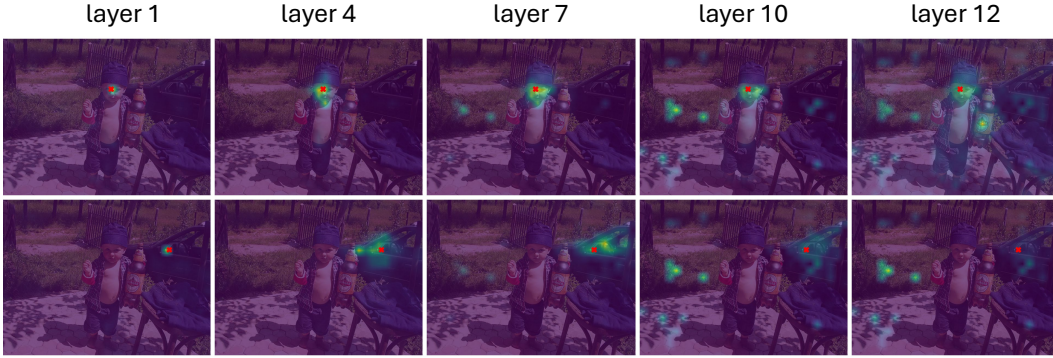


Figure 2: Visualization of the abnormal token phenomenon in the attention maps across different layers of the ViT-B/16 model in the CLIP vision encoder.

138 We present the layerwise SD and SA scores for both the ViT-B/16 and ViT-L/14 models used as the
 139 CLIP vision encoder in Figure 1. From the results, we make the following observations:

140 • The SD of CLIP exhibits an inverted U-shaped curve across layers: it initially increases in the early
 141 stages but declines in the deeper layers. This decline is especially prominent in the final layers. For
 142 example, the last two layers ((excluding the final layer)) of the ViT-B/16 model and the last five
 143 layers of the ViT-L/14 model show a marked reduction in spatial discriminability.

144 • SA follows an approximately monotonic increasing pattern across layers: it improves substantially
 145 in the early layers but gradually saturates in the final layers, offering only marginal gains thereafter.

146 These findings offer a nuanced understanding of why CLIP has proven effective for open-vocabulary
 147 semantic segmentation. In particular, the strong semantic alignment observed in the final layers
 148 explains why prior work often leverages last-layer features for aligning visual tokens with textual
 149 categories. However, the significant decline in spatial discriminability in these layers reveals a key
 150 limitation as they may lack the fine-grained spatial distinctions necessary for producing accurate and
 151 precise segmentation masks. In this work, we aim to address this limitation by proposing methods
 152 that jointly preserve spatial structure and semantic alignment through a systematic exploitation of
 153 spatial discriminability across token, layer, and head levels. Before introducing our approach, we first
 154 investigate the underlying causes of the decline in spatial discriminability in the next subsection.

2.3 Analysis of abnormal tokens

Class-agnostic sparse and large-norm tokens. To understand the progression within the vision encoder, we analyze attention maps across layers. As shown in Figure 2, deeper layers exhibit a small set of dominant tokens that receive disproportionately high attention from nearly all spatial locations, causing most tokens to focus on this subset, consistent with prior observations [33, 18]. This leads to a gradual decline in spatial discriminability, which is essential for accurate segmentation. To further characterize these dominant tokens, we compare their features with those of normal tokens. As illustrated in Figure 3, dominant tokens exhibit sparse and consistent activation patterns, with only a few channels maintaining high activation. To quantify this sparsity, we adopt the hoyer score [34]:

$$\mathcal{H}(\mathbf{x}_i^l) = \frac{\sqrt{D} - \frac{|\mathbf{x}_i^l|_1}{|\mathbf{x}_i^l|_2}}{\sqrt{D} - 1}, \quad (4)$$

where $\mathbf{x}_i^l \in \mathbb{R}^D$ is the feature vector of the i -th token at layer l . We use this metric to quantify sparsity and visualize its distribution across layers and token positions in Figure 3(b). To evaluate whether dominant tokens encode meaningful semantics, we analyze their pairwise cosine similarity across spatial locations, layers, and image samples on the ImageNet validation set. As shown in Figure 4, these tokens exhibit strong invariance across positions and inputs, indicating limited semantic specificity. Contrary to prior assumptions that they capture global semantic content, our results suggest they act more like bias components that offset global-mean features, facilitating text alignment, similar to the bias term in final-layer classifiers under neural collapse [35, 36].

3 Method

In this section, we provide a detailed description of our training-free framework, which comprises three components: Abnormal Token Replacement (ATR) in Section 3.1, Spatial-Semantic Reweighting (SSR) in Section 3.2, and Selective Head Enhancement (SHE) in Section 3.3. Each component is complementary, and together they work synergistically to enhance the spatial discriminability of the CLIP model, based on our previous analysis.

3.1 Abnormal token replacement (ATR)

To mitigate the adverse effects of these anomalous tokens, we propose a simple yet effective strategy to suppress their influence prior to the final layer. As demonstrated in our earlier analysis, these tokens exhibit characteristically sparse activation patterns. To systematically identify them, we employ the hoyer score $\mathcal{H}(\mathbf{x}_i^l)$ defined before as a sparsity-based criterion. Tokens with scores exceeding a predefined threshold τ are deemed anomalous and grouped into the set $\mathcal{A}_l = \{i | \mathcal{H}(\mathbf{x}_i^l) > \tau\}$. After identifying them, we suppress their influence using an unnormalized 2-dimensional Gaussian kernel. Specifically, each anomalous token at spatial position $(m, n) \in \mathcal{A}$ is replaced by a weighted aggregation of its neighboring non-anomalous tokens:

$$\mathbf{x}_{m,n}^l = \frac{\sum_{i=1}^w \sum_{j=1}^h w_{m,n,i,j}^l \mathbf{x}_{i,j}^l}{\sum_{i=1}^w \sum_{j=1}^h w_{m,n,i,j}^l}, \quad \forall (m, n) \in \mathcal{A} \quad (5)$$

$$w_{m,n,i,j}^l = \begin{cases} 0, & \text{if } (i, j) \in \mathcal{A} \\ \exp\left(-\frac{(m-i)^2 + (n-j)^2}{2\sigma^2}\right), & \text{otherwise} \end{cases} \quad (6)$$

Here, σ controls the spatial extent of smoothing, and the weights $w_{m,n,i,j}$ ensure that only normal tokens contribute to the reconstruction of anomalous ones. Empirically, we find that applying this strategy before the penultimate layer leads to a performance drop, likely due to the removal of inherent biases encoded in abnormal tokens, which substantially alters the inference process. Therefore, we apply it only at the penultimate layer, i.e., with $l = L - 1$.

3.2 Spatial-semantic reweighting (SSR)

After mitigating the influence of anomalous tokens in the input to the last layer, the model exhibits improved spatial discriminability. However, a critical challenge remains: anomalous tokens present

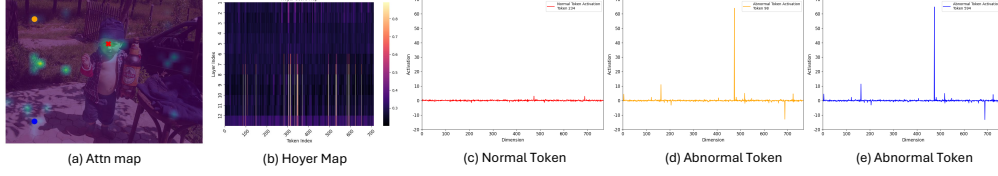


Figure 3: Illustration of the sparsity and high-norm characteristics of abnormal tokens. Figure (a) shows the attention map of the red anchor token. Figure (b) presents the Hoyer score distribution across layers and spatial positions. Figures (c)–(e) depict the channel activations of a normal token (red) and two abnormal tokens (yellow and blue) highlighted in Figure (a).

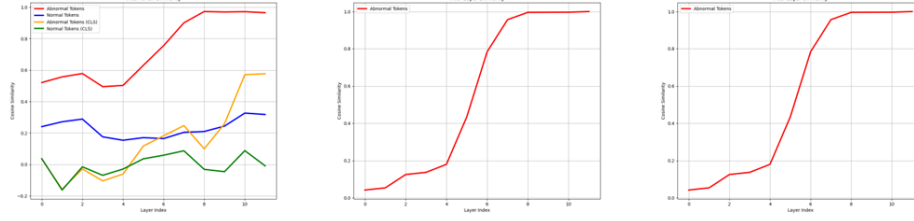


Figure 4: Layer-wise cosine similarity among abnormal tokens across positions, layers and samples.

in earlier layers may have already disrupted the spatial coherence of feature representations, limiting the effectiveness of final-layer refinements. Based on our layer-wise analysis, the final few layers overly emphasize alignment with text embeddings, the marginal gains in semantic alignment come at the cost of a pronounced decline in spatial discriminability. To address this imbalance, we propose a spatial-semantic reweighting strategy that enhances the model’s spatial awareness while preserving its semantic alignment capabilities. Given the feature representation \mathbf{X}^{l-1} at the l -th layer within the final few layers (e.g., layers 10–11 in ViT-B/16 and layers 20–23 in ViT-L/14), we reweight the forward pass by upweighting the residual pathway and downweighting the attention and MLP submodules, as follows:

$$\hat{\mathbf{X}}^l = (1 + \alpha)\mathbf{X}^{l-1} + (1 - \alpha)\text{Attn}(\text{LN}(\mathbf{X}^{l-1})), \quad (7)$$

$$\mathbf{X}^l = (1 + \alpha)\hat{\mathbf{X}}^l + (1 - \alpha)\text{FFN}(\text{LN}(\hat{\mathbf{X}}^l)), \quad (8)$$

where $\alpha \in [0, 1]$ is a reweighting coefficient that controls the relative degree of emphasis on the residual signal. As α increases, the l -th block increasingly preserves spatially discriminative features from earlier layers via the residual pathway, while diminishing the dominant influence of semantic aggregation in the attention and MLP submodules. To the best of our knowledge, prior work has primarily focused on reforming the final layer or modifying its representations to improve performance. However, these approaches often inherit the global semantic alignment bias inherent in the final few layers, resulting in a substantial decline in the spatial discriminability of the extracted features. In contrast, our SSR strategy explicitly mitigates this limitation by rebalancing the contributions of residual and semantic components in intermediate layers preceding the final layer.

3.3 Selective head enhancement (SHE)

Strong spatial discriminability of some attention heads. While the proposed strategies effectively enhance the spatial discriminability in the final layers, the overall spatial discriminability of the features output by the CLIP vision encoder may still remain suboptimal. Inspired by recent studies [37, 38] revealing that different attention heads capture distinct visual concepts, such as number, shape and texture, this motivates us to investigate whether certain heads are specifically responsible for encoding spatial discriminability. To identify such heads, we follow the formulation introduced in [39, 37], which rewrites the multi-head self-attention (MSA) output as a summation over H independent attention heads: $\text{Attn}(\text{LN}(\mathbf{X}^l)) = \sum_{h=1}^H \mathbf{A}_h^l \mathbf{V}_h^l \mathbf{W}_o^l \in \mathbb{R}^{(1+hw) \times D}$, where \mathbf{A}_h^l and \mathbf{V}_h^l denote the attention and value matrices for the h -th head at layer l , and \mathbf{W}_o^l is the output projection matrix shared across all heads. We extract the contribution of the h -th head at layer l and apply abnormal token resolution as follows:

$$\mathbf{X}^{l,h} = \sigma(\mathbf{A}_h^l \mathbf{V}_h^l \mathbf{W}_o^l), \quad (9)$$

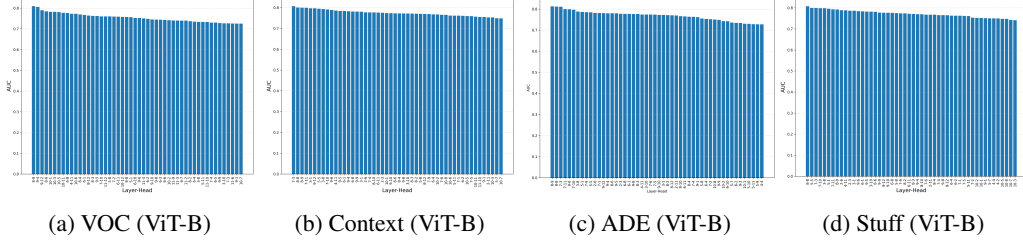


Figure 5: Head-wise analysis of spatial discriminability within the ViT-B/16 vision backbone across multiple datasets. To ensure consistency, the final layer is excluded, and only the top 50 attention heads are visualized in each figure for clarity.

where $\sigma(\cdot)$ denotes the abnormal token replacement operation defined previously. To assess the SD of each attention head, we adopt the same AUC-based metric as the preceding layer-level analysis. Figure 5 shows the head AUC distribution for ViT-B/16, with ViT-L/14 results in the appendix. From the figure, we observe that the output features from certain attention heads, such as the 9th head in the 8th layer, consistently exhibit high AUC scores across different datasets, suggesting that these heads are more effective at capturing SD information than others.

Building on this observation, we propose to selectively leverage high-performing heads to enhance the spatial discriminability of the output representations. Let $AUC_{l,h}^s$ denote the AUC score of the representations from the h -th head in the l -th layer for dataset $s \in \{\text{VOC, Context, ADE, Stuff}\}$. To obtain a dataset-agnostic measure of discriminability, we compute the average AUC score for each head across all datasets, denoted as $\overline{AUC}_{l,h}$. The distribution of these average scores is provided in the appendix. We rank all heads by their $\overline{AUC}_{l,h}$ scores and select the top- k to form the set $\mathcal{H}_{\text{top}k}$. The corresponding feature representations are then aggregated as: $\overline{\mathbf{X}}_{\text{top}k} = \frac{1}{k} \sum_{(l,h) \in \mathcal{H}_{\text{top}k}} \mathbf{X}^{l,h}$. This aggregated feature $\overline{\mathbf{X}}_{\text{top}k}$ is used to construct a similarity map $S = \frac{\overline{\mathbf{X}}_{\text{top}k} \overline{\mathbf{X}}_{\text{top}k}^\top}{\|\overline{\mathbf{X}}_{\text{top}k}\|^2}$, which captures the pairwise similarity among visual tokens. To mitigate the influence of spurious interactions between tokens from different semantic categories, we apply a thresholding operation with a predefined parameter β , resulting in the filtered similarity map S_β , where $S_\beta(i, j) = S(i, j)$ if $S(i, j) \geq \beta$, and $S_\beta(i, j) = 0$ otherwise. The resulting S_β is then column-wise normalized, and subsequently used to refine the final-layer features by $\mathbf{X}^{L-1} = \text{Norm}(S_\beta) \mathbf{X}^{L-1}$.

4 Experiment Results

Evaluation datasets. We follow the standard evaluation protocol from prior works [3, 15, 16] and assess our method on eight widely used semantic segmentation benchmarks. For clarity, we group them into two categories and use abbreviated names throughout the paper. The first category excludes background and includes Pascal VOC [29] (VOC20), Pascal Context [30] (Context59), COCO-Stuff [32] (Stuff), ADE20K [31] (ADE), and Cityscapes [40] (City). The second includes background and consists of VOC21, Context60, and COCO-Object [32] (Object). We use CLIP [1] models with ViT-B/16 and ViT-L/14 backbones via MMSegmentation [41], and report results using the mean Intersection-over-Union (mIoU). All hyperparameters are fixed across datasets without task-specific tuning. Additional implementation details are provided in the appendix.

4.1 Comparison with existing methods.

We compare our approach against a comprehensive set of open-vocabulary semantic segmentation (OVSS) methods, including the direct baseline CLIP [1], as well as several state-of-the-art training-free approaches: MaskCLIP [14], CLIPSurgery [13], SCLIP [3], NACLIP [16], ClearCLIP [15], LAVG [42], and ResCLIP [20]. We also include several influential weakly supervised methods, such as GroupViT [5], ReCo [43], and TCL [8]. Unless otherwise specified, all reported results are taken directly from the respective original papers and ResCLIP [20]. As our method is orthogonal to approaches that primarily target improvements in the final-layer attention, we additionally evaluate its effectiveness when integrated with recent state-of-the-art methods that employ specialized attention mechanisms in the last layer, including SCLIP [3], ClearCLIP [15], and ResCLIP [20]. For fair

Table 1: Performance comparison of our approach with other methods on eight semantic segmentation benchmarks following the evaluation protocol in Section 4. Our results are marked in gray.

Methods	Training	With a background class			Without background class				Avg.
		VOC21	Context60	Object	VOC20	City	Context59	ADE Stuff	
ReCo [43]	✓	25.1	19.9	15.7	57.7	21.1	22.3	11.2 14.8	23.5
GroupViT [5]	✓	52.3	18.7	27.5	79.7	18.5	23.4	10.4 15.3	30.7
TCL [8]	✓	51.2	24.3	30.4	77.5	23.1	30.3	14.9 19.6	33.9
CLIP [1]	✗	16.2	7.7	5.5	41.8	5.5	9.2	2.1 4.4	11.6
MaskCLIP [14]	✗	38.8	23.6	20.6	74.9	16.4	26.4	9.8 14.8	28.2
CLIPSurgery [13]	✗	55.2	18.7	27.5	79.7	18.5	23.4	10.4 15.3	31.1
LaVG [42]	✗	62.1	31.6	34.2	82.5	26.2	34.7	15.8 23.2	38.8
NACLIP [16]	✗	58.9	32.2	33.2	79.7	35.5	35.2	17.4 23.3	39.4
SCLIP [3]	✗	59.7	31.7	33.5	81.5	32.3	34.5	16.5 22.7	39.1
+TLH-CLIP (ours)	✗	64.8	34.8	36.6	86.3	36.1	37.6	18.0 24.9	42.4 (+3.3)
ClearCLIP [15]	✗	57.0	32.2	32.5	82.3	32.8	35.8	17.3 24.0	39.2
+TLH-CLIP (ours)	✗	63.9	35.2	35.6	85.7	37.8	38.8	19.2 25.8	42.7 (+3.5)
ResCLIP [20]	✗	60.0	32.7	34.0	85.5	35.6	35.8	17.7 23.8	40.6
+TLH-CLIP (ours)	✗	63.9	35.5	35.3	86.8	38.2	38.2	19.1 25.5	42.8 (+2.2)

comparison, we exclude the *Semantic Feedback Refinement* module in ResCLIP, as it relies on the computationally expensive PAMR [44] post-processing, which is inconsistent with our evaluation setting. For comprehensiveness, results on the ViT-L/14 architecture are provided in the appendix.

In Table 1, we summarize the performance of various open-vocabulary semantic segmentation models on benchmark datasets using the ViT-B/16 backbone. Our proposed TLH-CLIP consistently enhances the performance of state-of-the-art approaches, including SCLIP [3], ClearCLIP [15], and ResCLIP [20]. Notably, when integrated with ResCLIP [20], TLH-CLIP achieves state-of-the-art results, outperforming leading weakly supervised methods. As a plug-and-play solution, TLH-CLIP yields consistent improvements across all datasets compared to the respective baselines, demonstrating its strong generalization capability. We further evaluate performance on the ViT-L/14 backbone. In line with observations from [20], existing methods generally exhibit a performance drop exceeding 2% mIoU when adapting to a different backbone; for instance, ClearCLIP [15] suffers a notable decline of 2.7% mIoU. In contrast, when augmented with TLH-CLIP, this performance degradation is significantly alleviated, highlighting the robustness of our approach. Across both backbones, TLH-CLIP delivers substantial improvements over baseline methods, validating its effectiveness.

4.2 Experimental analysis

In this section, we conduct comprehensive ablation studies to validate the effectiveness of our proposed method. We adopt SCLIP [3] as the baseline, which enhances spatial correlation by modifying the attention mechanism in the final layer, replacing the standard QK^\top attention with a combination of $QQ^\top + KK^\top$. In addition, following prior work [15, 20], we remove the residual connections and FFN from the final transformer block to isolate the impact of attention refinement.

Analysis of the hoyer threshold parameter τ . Our method relies on hoyer sparsity to identify anomalous tokens, making the sparsity threshold τ a critical hyperparameter. We conduct a systematic evaluation, as shown in Table 2. At $\tau = 0.2$, many normal tokens are misclassified, leading to excessive smoothing and degraded performance. As τ increases to 0.4, performance steadily improves, but plateaus between 0.5 and 0.8, with a decline observed beyond this range. The broad stable region indicates a clear sparsity gap between normal and abnormal tokens, highlighting the robustness of ATR to threshold selection. Based on this analysis, we fix $\tau = 0.5$ for all experiments.

Analysis of spatial-semantic reweighting parameters and number of layers. To evaluate the impact of the reweighting strength α and the range of layers involved, from l_{start} to l_{end} , we perform a comprehensive sensitivity analysis. The results are summarized in Table 3. We observe that the best performance is obtained when reweighting is applied to layers 10–11 in the ViT-B/16 backbone. This aligns with our earlier findings that these layers experience a marked decline in spatial discriminability while yielding only marginal improvements in semantic alignment. Extending reweighting to include layer 9 results in a slight gain in spatial discriminability but introduces noisy

Table 2: Study of hoeyer sparsity threshold τ .

τ	C60	Obj	C59	City	Avg
$\tau = 0.2$	0.8	2.0	1.5	1.7	1.5
$\tau = 0.4$	32.8	34.0	36.6	34.7	34.5
$\tau = 0.5$	32.8	34.2	36.7	34.7	34.6
$\tau = 0.8$	32.8	33.9	36.6	34.7	34.5
$\tau = 0.9$	32.8	33.9	36.6	34.3	34.4
baseline	32.4	32.9	36.0	34.3	33.9

Table 3: Study of $(l_{\text{start}}, l_{\text{end}}, \alpha)$ in SSR module.

$(l_{\text{start}}, l_{\text{end}}, \alpha)$	C60	Obj	C59	City	Avg
baseline	32.4	32.9	36.0	34.3	33.9
(9, 11, 0.1)	32.7	32.0	36.5	36.7	34.5
(10, 11, 0.1)	33.1	33.4	36.9	35.6	34.8
(11, 11, 0.1)	32.7	34.1	36.4	34.9	34.5
(10, 11, 0.05)	32.8	33.7	36.4	35.0	34.5
(10, 11, 0.2)	32.6	31.7	36.5	36.6	34.4

Table 4: Study of number of selected heads k .

k	C60	Obj	C59	City	Avg
baseline	32.8	34.2	36.7	34.7	34.6
layer($l = 8$)	33.9	37.1	37.1	35.0	35.8
$k = 1$	33.4	37.1	36.6	35.4	35.3
$k = 10$	34.8	37.6	37.9	36.3	36.7
$k = 30$	34.7	37.3	37.9	36.4	36.6
$k = 50$	34.7	37.3	37.8	36.3	36.5

Table 5: Combination of three strategies.

Methods	Module			mIoU	Δ
	ATR	SSR	SHE		
baseline	—	—	—	33.9	—
	✓	✓	—	35.3	+1.4
	✓	—	✓	36.7	+2.8
Ours	✓	✓	✓	37.4	+3.5

semantic signals, ultimately leading to a reduction in segmentation performance. In addition, we examine the effect of varying the reweighting threshold parameter α . As α increases from 0 to 0.1, performance improves steadily, indicating a beneficial balance between spatial and semantic cues. However, further increasing α leads to a performance drop, as it incorporates more noisy semantic information from earlier layers and significantly perturbs the input distribution of subsequent layers.

Analysis of the number of selected heads. We study the effect of varying the number of top- k attention heads selected for enhancement, as shown in Table 4. Empirically, we find that SHE is most effective when combined with ATR; without ATR, the spatially coherent similarity maps can cause normal tokens to be fused with abnormal ones. Therefore, we adopt the baseline SCLIP model equipped with ATR as our baseline. On the ViT-B/16 backbone, increasing k from 1 to 10 improves segmentation accuracy, as aggregating multiple spatially discriminative heads helps suppress spurious correlations. However, performance declines when k becomes too large due to the inclusion of noisy or less informative heads, which introduce undesired cross-category interactions. We also compare head- and layer-level selection (best $l = 8$), finding that head-level selection consistently performs better, as discriminative heads are distributed across layers, while entire-layer selection introduces irrelevant heads and degrades performance.

Study of each individual components In the previous parts, we evaluated the effectiveness of each individual component. Table 5 presents their combinations, which yield a substantial improvement of 3.5 mIoU, achieving a final mIoU of 37.5 on these four datasets. These results highlight the complementary contributions of each module to the overall segmentation performance.

5 Conclusion

In this paper, we present a comprehensive analysis of the spatial discriminability of pretrained CLIP models across the token, layer, and head levels. Our study reveals three key findings: (1) the emergence of class-agnostic abnormal tokens with sparse, high-norm activations; (2) a notable decline in spatial discriminability in the final layers, despite marginal gains in semantic alignment; and (3) consistently strong spatial discriminability in specific attention heads. Motivated by these observations, we propose TLH-CLIP, a training-free framework that enhances spatial discriminability while preserving semantic alignment. TLH-CLIP introduces three complementary components: (1) abnormal token replacement, (2) spatial-semantic reweighting, and (3) selective head enhancement. Unlike prior methods that focus on modifying the final attention layer, our approach provides lightweight, plug-and-play modules compatible with existing architectures. Extensive experiments on multiple segmentation benchmarks demonstrate that TLH-CLIP consistently outperforms strong baselines. Moreover, as CLIP vision encoders are often frozen during the training of MLLMs, our findings offer valuable insights for improving visual understanding in broader MLLMs.

References

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [2] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.
- [3] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024.
- [4] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.
- [5] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022.
- [6] Yun Xing, Jian Kang, Aoran Xiao, Jiahao Nie, Ling Shao, and Shijian Lu. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36:68798–68809, 2023.
- [7] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. *arXiv preprint arXiv:2412.04533*, 2024.
- [8] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023.
- [9] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.
- [10] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023.
- [11] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2935–2944, 2023.
- [12] Fei Zhang, Tianfei Zhou, Boyang Li, Hao He, Chaofan Ma, Tianjiao Zhang, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. *Advances in Neural Information Processing Systems*, 36:73652–73665, 2023.
- [13] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pages arXiv–2304, 2023.
- [14] 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.
- [15] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024.
- [16] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5061–5071. IEEE, 2025.
- [17] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024.
- [18] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 139–156. Springer, 2024.

- [19] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024.
- [20] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. *arXiv preprint arXiv:2411.15851*, 2024.
- [21] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 320–337. Springer, 2024.
- [22] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing correlations in clip with off-the-shelf foundation models for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.10086*, 2024.
- [23] Barbara Toniella Corradini, Mustafa Shukor, Paul Couairon, Guillaume Couairon, Franco Scarselli, and Matthieu Cord. Freeseg-diff: Training-free open-vocabulary segmentation with diffusion models. *ArXiv*, abs/2403.20105, 2024.
- [24] Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.13836*, 2024.
- [25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [28] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023.
- [29] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep.*, 8(5):2–5, 2011.
- [30] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.
- [31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.
- [32] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.
- [33] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.
- [34] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.
- [35] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.
- [36] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pages 27179–27202. PMLR, 2022.
- [37] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip’s image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

- [38] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. *arXiv preprint arXiv:2503.06287*, 2025.
- [39] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.
- [40] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [41] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mms Segmentation>, 2020.
- [42] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 143–164. Springer, 2024.
- [43] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022.
- [44] Nikita Araslanov 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.
- [45] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [46] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.
- [47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
- [48] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.
- [49] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23681–23690, 2024.
- [50] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. V1-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022.
- [51] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023.
- [52] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.
- [53] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yanan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, June 2023.
- [54] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, 2022.

- 484 [55] Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, and Boyang Albert Li. Emergent open-vocabulary
 485 semantic segmentation from off-the-shelf vision-language models. *2024 IEEE/CVF Conference on*
 486 *Computer Vision and Pattern Recognition (CVPR)*, pages 4029–4040, 2023.
- 487 [56] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer.
 488 *ArXiv*, abs/2206.07045, 2022.
- 489 [57] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-
 490 vocabulary semantic segmentation models from natural language supervision. *2023 IEEE/CVF Conference*
 491 *on Computer Vision and Pattern Recognition (CVPR)*, pages 2935–2944, 2023.
- 492 [58] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-
 493 vocabulary panoptic segmentation with text-to-image diffusion models. *2023 IEEE/CVF Conference on*
 494 *Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023.
- 495 [59] Reza Qorbani, Gianluca Villani, Theodoros Panagiotakopoulos, Marc Botet Colomer, Linus Harenstam-
 496 Nielsen, Mattia Segu, Pier Luigi Dovesi, Jussi Karlgren, Daniel Cremers, Federico Tombari, and Matteo
 497 Poggi. Semantic library adaptation: Lora retrieval and fusion for open-vocabulary semantic segmentation.
 498 *ArXiv*, abs/2503.21780, 2025.
- 499 [60] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge
 500 potentials. *Advances in neural information processing systems*, 24, 2011.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly state the claims about of analysis discovery and the proposed method, matching our experimental results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: As noted in the paper, the proposed methods can mitigate but not entirely resolve the decline in spatial discriminability in the final layers.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed descriptions of the experimental settings in Section 4 and the appendix. Additionally, the ablation studies present the rationale behind the choice of hyperparameters used in this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer:[Yes]

Justification: The implementation code is included in the supplementary materials and will be made publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide detailed descriptions of the experimental settings in Section 4 and the appendix. Additionally, the ablation studies present the rationale behind the choice of hyperparameters used in this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: Since our method is training-free and directly uses the pretrained CLIP model weights without any additional optimization, issues related to statistical significance do not arise.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: The details about computer resources used in the experiments are reported in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The paper adheres to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: he paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original owners of assets (e.g., code, data, models), used in the paper are properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Technical Appendices and Supplementary Material

In the appendix, we first provide additional related literature in Appendix A.1. We then present extended experimental details and results for both the ViT-B/16 and ViT-L/14 vision encoders in Appendix A.2. For each model, we report the distribution of average head AUC scores across multiple datasets, conduct a component-wise analysis of their contributions to spatial discriminability, and provide qualitative comparisons along with additional supporting results and visualizations.

A.1 Additional literature

Vision-language pre-training models. Deep learning is experiencing a significant paradigm shift driven with the emergence of large-scale vision-language models. Among these, the Contrastive Language-Image Pre-Training (CLIP) framework [1] has achieved remarkable success, largely attributed to its strong zero-shot and few-shot generalization capabilities in visual recognition tasks, particularly image classification. Building on these strengths, an expanding body of research has sought to further improve CLIP’s zero-shot performance through enhanced training on large-scale image-text pairs [45, 46, 47], or by enabling its efficient adaptation to novel downstream tasks using limited labeled data [48, 49, 50, 51, 52]. Nevertheless, as CLIP is pre-trained predominantly at the image level, its learned representations, particularly the [CLS] token, are optimized to capture global semantics. This coarse-grained supervision inherently limits their utility in dense prediction tasks, where fine-grained spatial localization is critical.

Open-vocabulary Semantic Segmentation. Driven by the remarkable generalization capabilities of large-scale vision-language models [1, 45, 46, 47], a growing line of research has focused on open-vocabulary semantic segmentation [10, 53, 54, 55, 56, 57, 58, 7, 59], which seeks to extend global cross-modal alignment to fine-grained, pixel-level predictions. Existing approaches can be broadly categorized into three groups: *fully supervised*, *weakly supervised*, and *training-free* methods, based on the level of auxiliary supervision required for adaptation. *Fully-supervised* approaches [53, 55, 7, 59] adapt pretrained CLIP models to semantic segmentation by leveraging large-scale datasets with dense pixel-wise annotations for a predefined, yet limited, set of categories. Although such supervision facilitates the learning of fine-grained spatial representations, these methods often struggle to generalize to novel categories. Furthermore, their reliance on labor-intensive, densely annotated datasets poses significant challenges to scalability, limiting their applicability in real-world scenarios. Instead of requiring pixel-wise annotations, *weakly-supervised* approaches leverage auxiliary datasets with image-level annotations to adapt pre-trained CLIP models for semantic segmentation. These approaches commonly employ large-scale image-text corpora, wherein textual descriptions explicitly reference the object categories present in the corresponding images. Adaptation is achieved by enforcing cross-modal alignment through a contrastive loss [5], analogous to the original CLIP pre-training objective. Although such strategies alleviate the burden of dense supervision, they still rely on access to large-scale annotated datasets. However, the auxiliary datasets are substantially smaller than those employed during CLIP pre-training, which limits the generalization capability. Additionally, these methods presuppose prior knowledge of the categories present in each image, thereby constraining their applicability in genuinely open-world segmentation scenarios.

Training-free open-vocabulary semantic segmentation. *Training-free* methods explore the feasibility of employing frozen CLIP models to generate segmentation masks in the absence of additional data for adaptation. Some approaches rely on auxiliary models pre-trained on large-scale datasets such as DINO [21, 17], SAM [22], or Stable Diffusion [23, 24], which incur significant computational and memory overhead. An alternative line of research seeks to enhance CLIP’s capability for dense visual representation by modifying the inference pipeline of its visual encoder. For instance, MaskCLIP [14] removes the self-attention module in the final Transformer layer and demonstrates that the resulting value embeddings encode local visual features that align effectively with textual prompts. CLIP-Surgery [13] introduces a dual-path structure that replaces the original self-attention with value-value attention to better preserve semantic consistency, mitigating the tendency to attend to unrelated regions. Building on this idea, GEM [19] and SCLIP [3] generalize the approach by incorporating correlative self-attention mechanisms, such as query-query and key-key interactions. ClearCLIP [15] further simplifies the architecture by removing the final feed-forward layer and associated residual connections, which were found to contribute to noisy segmentation outputs. Addi-

tionally, NACLIP [16] imposes explicit spatial regularization, encouraging each token to primarily attend to its neighbors to improve the spatial coherence.

Taking into account both generalization ability and computational costs, this work aims to advance existing training-free methods for open-vocabulary semantic segmentation. While prior training-free approaches predominantly rely on features extracted from the final layer of the visual encoder, we instead conduct a comprehensive analysis across token, head, and layer levels. Building on this multi-level analysis, we propose recipe at each level to enhance the spatial representation capacity of CLIP, thereby improving segmentation performance within a training-free framework.

A.2 Additional experimental results

A.2.1 Additional Implementation Details.

We adopt the CLIP [1] models with ViT-B/16 and ViT-L/14 backbones, implemented via the MM-Segmentation framework [41]. Following the protocol in [20], input images are resized such that the shorter side is 336 pixels, except for Cityscapes, which is resized to 560 pixels to accommodate its higher resolution. Inference is performed using a sliding-window strategy with a crop size of 224×224 and a stride of 112 pixels. Consistent with TCL [8], we refrain from employing computationally intensive post-processing techniques that may lead to unfair comparisons, such as PAMR [44] (used in TCL [8], NACLIP [16]) and DenseCRF [60] (used in ReCo [43]). For textual inputs, we utilize the standard ImageNet prompts [1] without incorporating any additional prompting strategies. Based on the ablation study in Section 4, we configure the model as follows:

- **ViT-B/16:** The hoyr threshold is set to $\tau = 0.5$, and the standard deviation of the Gaussian kernel is 0.5. The reweighting operation is applied to layers [10, 11] (i.e., the two layers preceding the final one), with a reweighting coefficient of 0.1. The top-10 attention heads with the highest spatial discriminability are selected using a filter threshold of $\beta = 0.7$, corresponding to the following (layer, head) pairs: (8,9), (8,8), (7,10), (9,12), (7,3), (9,4), (5,1), (9,6), (4,11), and (8,6).
- **ViT-L/14:** The hoyr threshold is set to $\tau = 0.4$, with the same Gaussian kernel standard deviation of 0.5. Reweighting is applied to layers [17, 23] with a coefficient of 0.1. Top-performing heads are selected using the same threshold $\beta = 0.7$, with the top-30 (layer, head) pairs identified as: (11, 3), (9, 3), (7, 9), (11, 6), (10, 10), (9, 13), (3, 10), (4, 14), (10, 6), (6, 9), (7, 12), (14, 16), (11, 8), (10, 13), (8, 4), (8, 8), (10, 8), (9, 4), (2, 11), (9, 6), (8, 1), (14, 1), (16, 2), (4, 13), (13, 11), (11, 14), (7, 4), (14, 11), (13, 13), and (3, 13).

All experiments are conducted using eight NVIDIA RTX A5000 GPUs, each with 24 GB of memory.

A.2.2 Additional analysis of individual components effects on spatial discriminability

The ablation study in Section 4 highlights the effectiveness of each individual strategy. To further clarify the contribution of each component, we conduct an incremental analysis that evaluates their impact on spatial discriminability. Specifically, we examine how the progressive integration of these strategies influences the quality of the final representation. As shown in Figure 6, we plot the ROC curves of the output features from the penultimate layer, where a higher area under the curve (AUC) signifies stronger spatial discriminability. Each individual strategy contributes to a consistent improvement in AUC. When combined, these strategies lead to a substantial enhancement in spatial discriminability. For the ViT-B/16 backbone, AUC scores improve from 0.7560 to 0.8320 on the VOC dataset, from 0.7406 to 0.8270 on the Context dataset, from 0.7205 to 0.8282 on the ADE dataset, and from 0.7278 to 0.8175 on the COCO-Stuff dataset. Similarly, for the ViT-L/14 backbone, the AUC increases from 0.7081 to 0.8460 on VOC, from 0.6574 to 0.8126 on Context, from 0.6383 to 0.8208 on ADE, and from 0.6447 to 0.8174 on COCO-Stuff. These results demonstrate the complementary nature of the proposed strategies and their collective effectiveness in enhancing spatial discriminability across diverse segmentation benchmarks.

Since our SSR strategy is applied not only to the penultimate layer but also to earlier layers, we further present the layer-wise curves of spatial discriminability and semantic alignment after applying SSR in Figure 7. As illustrated in the figure, the spatial discriminability of the final few layers is significantly enhanced, demonstrating the effectiveness of the proposed SSR strategy. Correspondingly, semantic alignment also exhibits consistent improvements, as the enhanced spatial structure contributes to better semantic consistency.

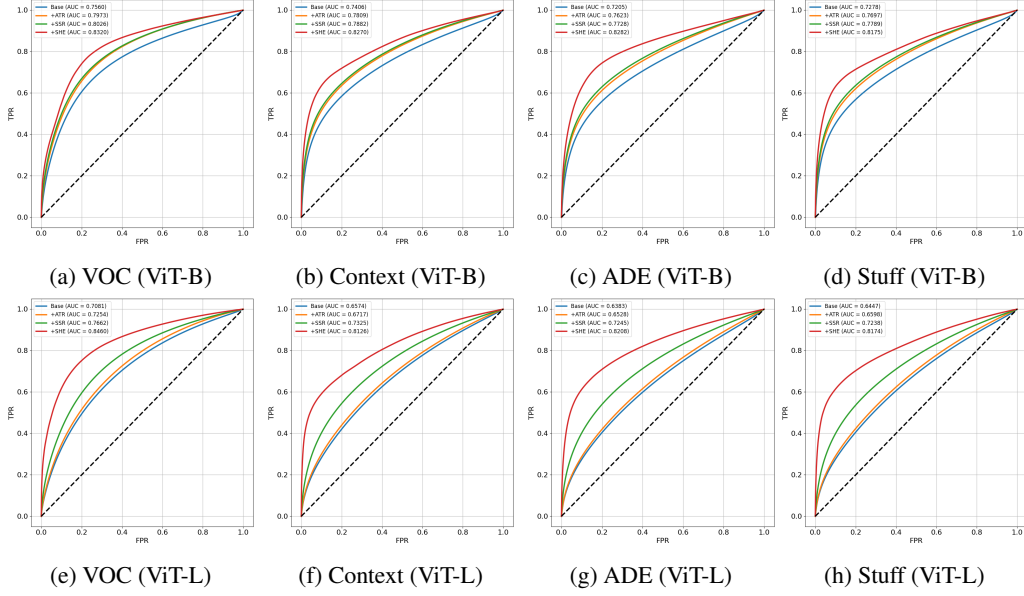


Figure 6: ROC curves of the penultimate-layer output features for ViT-B/16 and ViT-L/14 backbones across four datasets (VOC, Context, ADE, and COCO-Stuff). Each curve corresponds to a different combination of spatial discriminability enhancement strategies. The results demonstrate consistent improvements in AUC with each added strategy, culminating in significantly higher spatial discriminability when all components are applied.

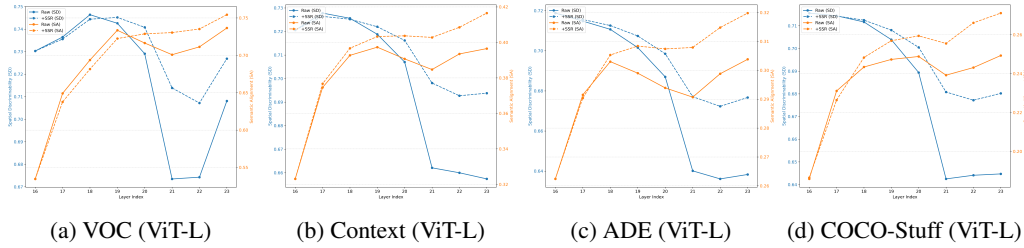


Figure 7: Layer-wise spatial discriminability (SD, blue) and semantic alignment (SA, orange) for ViT-L/14 on four benchmarks: (a) VOC, (b) Context, (c) ADE, and (d) COCO-Stuff. Dashed lines denote the baseline CLIP (Raw) performance, while solid lines indicate results after applying the proposed SSR strategy. The application of SSR notably improves spatial discriminability in the final layers and consistently enhances semantic alignment across all datasets, demonstrating its effectiveness in preserving both local structure and semantic coherence.

964 A.2.3 Head AUC distributions

965 In Section 3, we presented the distribution of head-level AUC scores across different datasets using
 966 ViT-B/16 as the visual encoder. For completeness, we also provide the corresponding head-level
 967 AUC distribution for the ViT-L/14 model in Figure 8. Similar to the observations from ViT-B/16,
 968 we consistently find that certain heads exhibit significantly higher spatial discriminability than
 969 others—for example, the (layer, head) pairs (11, 3), (9, 3), and (7, 9). Since we select the top- k heads
 970 based on their average spatial discriminability across multiple datasets, we additionally present the
 971 distribution of average spatial discriminability scores for all heads in descending order in Figure 9 for
 972 ViT-B/16 and Figure 10 for ViT-L/14, to facilitate reference and reuse in future research.

973 A.2.4 Additional experiments on ViT-L/14

974 In Table 6, we provide a detailed comparison of open-vocabulary semantic segmentation models
 975 on multiple benchmark datasets using the ViT-L/14 backbone. TLH-CLIP consistently improves
 976 the performance of leading methods, including ClearCLIP [15], NACLIP [16], and ResCLIP [20].

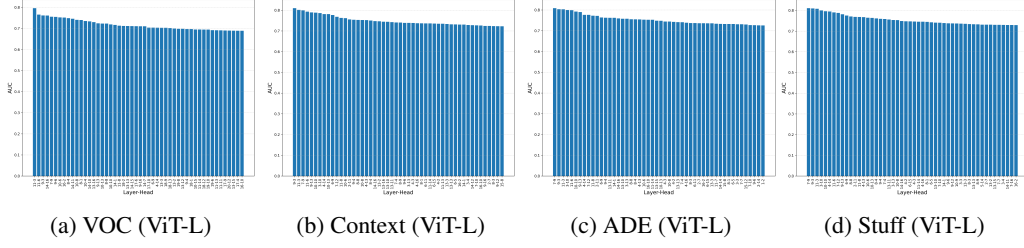


Figure 8: Head-wise analysis of spatial discriminability within the ViT-L/14 vision backbone across multiple datasets. To ensure consistency, the final layer is excluded, and only the top 50 attention heads are visualized in each figure for clarity.

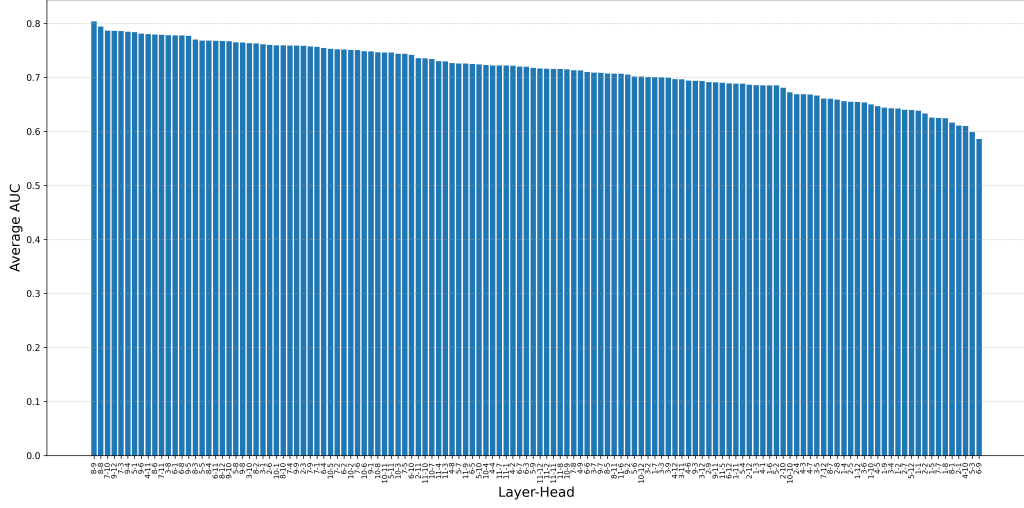
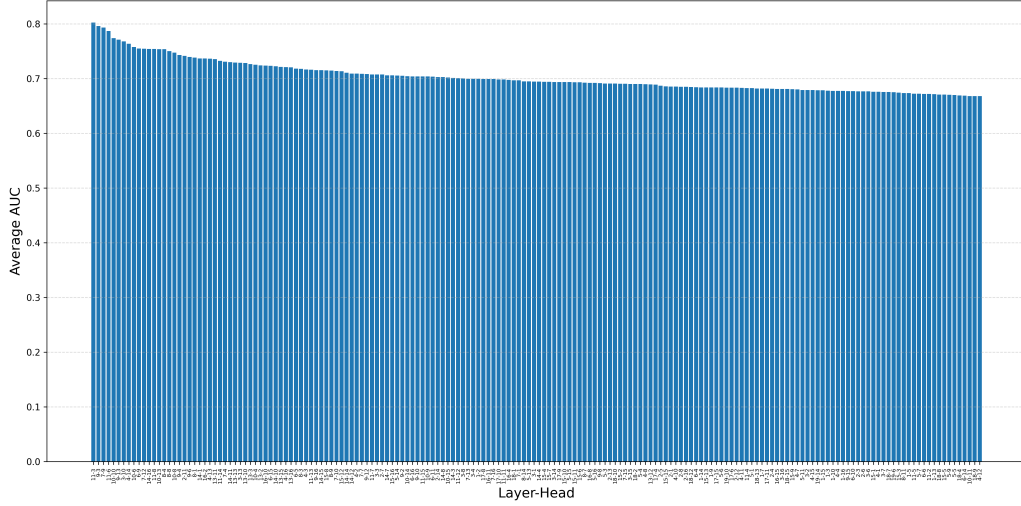


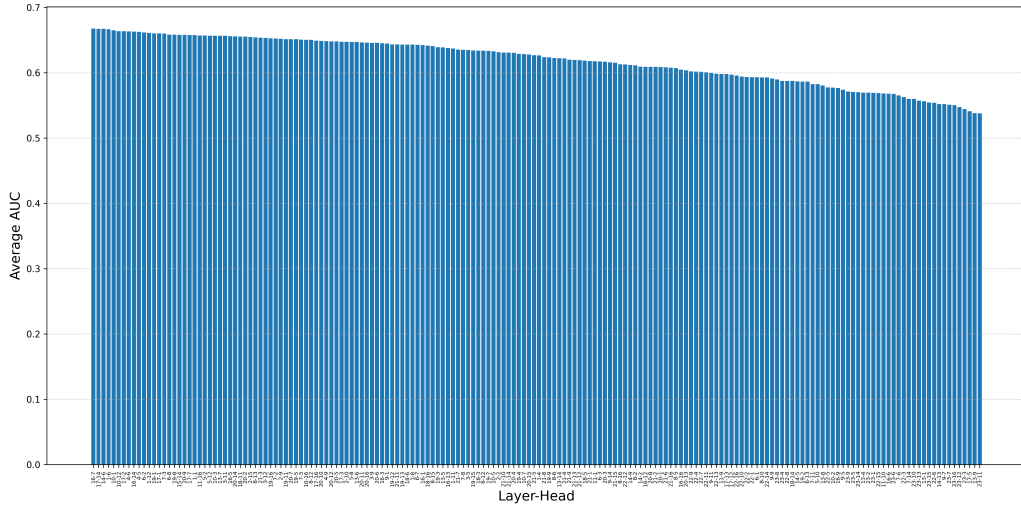
Figure 9: Distribution of average AUC scores for individual attention heads across four datasets using the ViT-B/16 model.

977 Notably, integrating TLH-CLIP with ResCLIP yields state-of-the-art results. As a plug-and-play
 978 module, TLH-CLIP delivers consistent improvements across all datasets, underscoring its robustness
 979 and generalization capability. However, we observe a noticeable performance drop when integrating
 980 TLH-CLIP into SCLIP [3]. This degradation may stem from the SSR component, which modifies
 981 a significant portion of the earlier transformer layers, potentially inducing distributional shifts in
 982 the inputs to the final-layer feed-forward network and compromising compatibility. As reported
 983 in [20], many methods suffer a performance drop exceeding 2% mIoU when adapted to a different
 984 backbone—for example, ClearCLIP experiences a 2.7% decline. In contrast, TLH-CLIP substantially
 985 mitigates such degradation, demonstrating strong cross-backbone robustness. These results further
 986 validate the effectiveness of TLH-CLIP in enhancing open-vocabulary segmentation performance
 987 across both ViT-B/16 and ViT-L/14 architectures.

988 Additionally, in Table 7, Table 8, Table 9, and Table 10, we present detailed studies on the hoier
 989 threshold parameter, the spatial-semantic reweighting configuration, the number of selected attention
 990 heads, and the ablation analysis of individual components, respectively, for the ViT-L/14 model.
 991 Consistent with the findings for the ViT-B/16 model, we observe that a moderate sparsity level
 992 with $\tau = 0.4$, a reweighting coefficient of $\alpha = 0.1$ applied to layers 17–23, and selecting the
 993 top- k heads with $k = 30$ yields the best overall performance. Accordingly, we adopt these settings
 994 as fixed hyperparameters for the ViT-L/14 backbone across all evaluated datasets. Moreover, the
 995 ablation study confirms the effectiveness of each individual component. As shown in Table 10, their
 996 combination leads to a substantial improvement of 4.7 mIoU, achieving a final score of 35.0 mIoU
 997 across the four benchmark datasets.



(a) First 176 heads



(b) Last 176 heads

Figure 10: Distribution of average AUC scores for individual attention heads across four datasets using the ViT-L/14 model.

A.2.5 Qualitative results

In Figure 11, we present a qualitative comparison between CLIP-based training-free baseline methods and their counterparts integrated with TLH-CLIP. As shown in the figure, the integration of TLH-CLIP consistently leads to more accurate and visually coherent segmentation results. The enhanced models produce cleaner segmentation maps with reduced noise, improved spatial consistency among same-category objects, and better delineation of object boundaries, all while preserving the original text-image alignment capabilities of the baseline methods.

A.2.6 Efficiency Comparison

In Table 11, we present an efficiency comparison on the Context59 dataset using an A5000 GPU with the ViT-B/16 backbone, selecting ClearCLIP as our baseline method. Compared to the previous state-of-the-art approach, ResCLIP, our method eliminates the need for the time-consuming PAMR post-processing step used in its Semantic Feedback Refinement module, while simultaneously improving segmentation performance. As a result, the inference speed increases from 3.0 to 8.2

Table 6: Performance comparison of our approach with other methods on eight semantic segmentation benchmarks following the evaluation protocol in Section 4. Our results are marked in gray.

Methods	Training	With a background class			Without background class					Avg.
		VOC21	Context60	Object	VOC20	City	Context59	ADE	Stuff	
CLIP [1]	\times	9.8	4.2	3.9	18.8	4.0	5.6	2.1	2.8	6.4
MaskCLIP [14]	\times	24.4	10.0	9.9	30.0	11.8	12.6	7.8	10.1	14.6
CLIPSurgery [13]	\times	47.7	27.2	27.5	80.5	32.1	31.5	16.4	17.3	35.0
SCLIP [3]	\times	44.4	22.3	24.9	60.3	32.2	20.5	7.1	13.1	28.1
ClearCLIP [15]	\times	48.7	28.3	29.7	80.0	27.9	29.6	15.0	19.9	34.9
+TLH-CLIP (ours)	\times	62.8	33.9	33.5	85.9	40.5	39.0	20.3	26.1	42.7 (+7.8)
NACLIP [16]	\times	52.2	28.7	30.4	78.7	31.4	32.1	17.3	21.4	36.5
+TLH-CLIP (ours)	\times	63.4	34.0	33.4	84.6	40.8	39.0	20.5	25.9	42.7 (+6.2)
ResCLIP [20]	\times	55.7	29.7	30.8	81.3	32.8	33.6	17.9	22.9	38.1
+TLH-CLIP (ours)	\times	63.4	34.0	33.4	85.9	40.9	39.0	20.5	26.0	42.9 (+4.8)

Table 7: Study of hoyer sparsity threshold τ .

τ	C60	Stuff	C59	City	Avg
$\tau = 0.2$	2.4	2.8	3.8	2.3	2.8
$\tau = 0.4$	29.7	22.5	33.6	34.8	30.2
$\tau = 0.5$	29.2	22.2	33.3	33.7	29.6
$\tau = 0.8$	28.9	22.0	33.0	33.2	29.3
$\tau = 0.9$	28.9	22.0	33.0	33.2	29.3
baseline	28.6	21.7	32.5	32.8	28.9

Table 8: Study of $(l_{\text{start}}, l_{\text{end}}, \alpha)$ in SSR module.

$(l_{\text{start}}, l_{\text{end}}, \alpha)$	C60	Stuff	C59	City	Avg
baseline	28.6	21.7	32.5	32.8	28.9
(14, 23, 0.1)	30.8	21.2	33.9	35.7	30.4
(17, 23, 0.1)	31.4	23.1	34.7	35.3	31.1
(19, 23, 0.1)	30.9	23.0	34.2	33.9	30.5
(17, 23, 0.05)	30.3	22.5	33.6	34.3	30.2
(17, 23, 0.2)	31.6	22.1	34.8	35.6	31.0

1011 FPS, and the computational cost is reduced from 141.34G to 102.65G FLOPs. Additionally, we also
1012 provide a detailed analysis of the computational overhead introduced by each individual strategy.

Table 9: Study of number of selected heads k .

k	C60	Stuff	C59	City	Avg
baseline	29.7	22.5	33.6	34.8	30.2
$k = 1$	33.5	25.3	36.9	37.3	33.3
$k = 10$	33.5	25.3	36.9	37.3	33.3
$k = 30$	33.6	25.4	37.1	37.8	33.5
$k = 60$	33.6	25.4	37.1	37.8	33.5
$k = 100$	33.5	25.4	37.1	37.7	33.4

Table 10: Combination of three strategies.

Methods	Module			mIoU	Δ
	ATR	SSR	SHE		
baseline	—	—	—	30.3	—
	✓	✓	—	32.8	+2.5
	✓	—	✓	33.5	+3.2
Ours	✓	✓	✓	35.0	+4.7

Table 11: Efficiency comparison of individual strategies.

Models	FLOPs(G) ↓	Params(M) ↓	Speed(FPS) ↑
CLIP	106.10	149.6	13.7
ResCLIP	141.34	149.6	3.0
baseline	100.70	149.6	13.9
+ATR	100.88	149.6	12.9
+SSR	100.88	149.6	11.7
+SHE	102.65	149.6	8.2

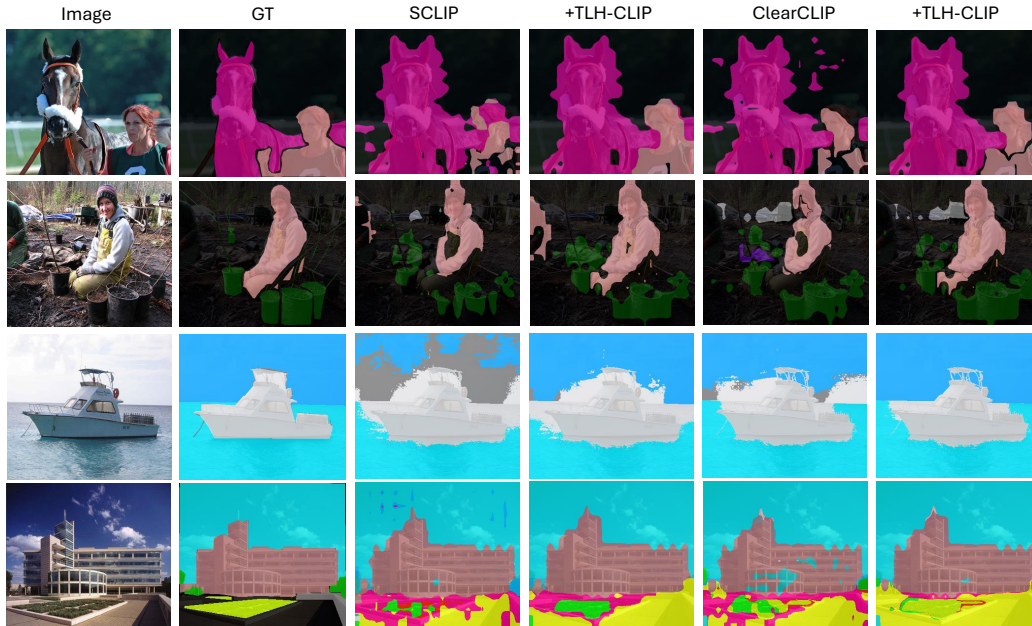


Figure 11: Qualitative comparison between CLIP-based training-free baseline methods and their counterparts integrated with TLH-CLIP.