CURE: Contextual Debiasing and Unbiased Refinement for Training-Free Open-Vocabulary Semantic Segmentation

Jiean Wang[1,2⋆], Sanqing Qu[1,2⋆], Fan Lu[1], Huanhuan Bao[3], Wei Tian[1], Bo Jin[1], Jiangtong Li[1], Junqiao Zhao[1], and Guang Chen[1,2†]

[1] Tongji University, Shanghai, China
[2] Shanghai Innovation Institution, Shanghai, China
[3] China Automotive Engineering Research Institute, Chongqing, China
{wangjiean2001, sanqingqu}@gmail.com, guangchen@tongji.edu.cn

Abstract. Recent advances in large-scale vision–language models (VLMs) have greatly advanced open-vocabulary semantic segmentation (OVSS), enabling segmentation beyond predefined category constraints. However, due to the image-level pre-training objectives of VLMs, existing training-free OVSS methods often suffer from global contextual artifacts, degrading their ability to capture fine-grained and spatially localized semantics. In this work, we propose a novel framework, termed CURE (Contextual debiasing and Unbiased REfinement), to address these limitations by mitigating both explicit and implicit global biases. Technically, CURE employs a reject-and-reconstruction module to identify explicit outlier tokens and reconstruct them using local neighborhood information. A global contextual debiasing module is further introduced to suppress residual implicit bias. Finally, a spatial correlation refinement module is designed to enhance spatial coherence by leveraging mid-level similarity patterns. Extensive experiments on five datasets demonstrate that CURE consistently improves segmentation quality. When integrated into six baseline methods, it yields average mIoU gains of 2.8% and 7.7% with ViT-B/16 and ViT-L/14 backbones, respectively.

Keywords: Open-vocabulary semantic segmentation · Feature debiasing · Visual-language model

1 Introduction

Semantic segmentation [2, 10, 28, 33, 48] aims to partition images into meaningful regions. Traditional methods, however, are constrained by predefined, fixed category labels [14,63], limiting their flexibility. As novel visual concepts continuously emerge, the demand for segmenting previously unseen categories grows [62, 63].

Open-vocabulary semantic segmentation (OVSS) has recently emerged as a promising paradigm that predicts pixel-level labels based on user-defined textual

⋆ Equal contribution.
† Corresponding author.

Fig. 1: Visualization of token attention patterns and segmentation outputs. The second column shows the average attention map across all tokens from an intermediate layer, revealing global context entanglement. The third column visualizes the final-layer CLS token attention, which encodes aggregated global information. The fourth column presents a norm-based heatmap highlighting outlier tokens that exhibit global attention artifacts. The fifth and sixth columns compare the segmentation results of ClearCLIP [22] with the refined outputs produced by our proposed method, CURE.

prompts, thus addressing limitations imposed by static category sets [11, 26, 51, 59]. Central to many OVSS methods are pre-trained vision-language models (VLMs) [19, 34, 56], which exhibit strong capabilities in aligning visual content with textual descriptions. This alignment enables effective handling of open-vocabulary tasks. However, the reliance on image-level alignment in such models often results in insufficient localization, limiting their performance in generating accurate pixel-wise predictions [51]. To alleviate these challenges, training-based approaches have been proposed for OVSS, which adopt VLMs through additional supervision to improve their performance on segmentation tasks [17, 20, 50, 51, 57, 61]. While effective, they typically require retraining on curated datasets and often rely on annotated examples from seen classes [8, 35, 49], constraining the scalability and generalization to novel concepts [11, 62]. In contrast, recent advances in training-free OVSS have leveraged pre-trained VLMs to improve segmentation without additional learning [9, 22, 23, 29], enhancing semantic associations within the image encoder to produce more accurate segmentation maps that generalize to unseen classes and adapt to dynamic scenarios.

Despite recent advances, existing methods remain limited by global artifacts induced by image-level pretraining. As shown in Figure 1, mid-level attention maps reveal that certain tokens inherently replicate the global attention pattern of the CLS token. This homogenization diminishes spatial discrimination and results in cluttered segmentation outputs. Such observations are consistent with prior analyses [5, 13], which have motivated modifications at the pretraining stage to alleviate global token dominance. Complementary to these efforts, we ask: *how can one directly suppress global contextual bias and recover spatially localized semantics without altering the pretrained model or introducing additional training?*

To this end, we propose CURE (Contextual debiasing and Unbiased REfinement), a novel training-free approach designed to counteract global contextual bias and enhance local feature saliency. Technically, CURE first employs a

Reject-and-Reconstruct module to eliminate outlier tokens that carry artifact-prone global patterns. It then applies a Global Contextual Debiasing mechanism to further suppress the influence of dominant global features embedded by pre-training. Finally, CURE leverages a Spatial Correlation Refinement module to refine the final-layer features by incorporating mid-level spatial correlations, which serve to recover fine-grained semantic structures often lost in biased representations. We extensively evaluate CURE on five benchmark datasets. Empirical results show that our method achieves new state-of-the-art performance.

Our contributions can be summarized as follows:

- We conduct a targeted analysis of token-level attention patterns in pretrained vision-language models and characterize how globally entangled tokens undermine spatial discrimination in training-free OVSS.
- We propose CURE, a novel training-free framework that mitigates global contextual bias and enhances the saliency of local features through a three-stage architecture. CURE is flexible and can be seamlessly incorporated into existing methods to further boost performance.
- We conduct comprehensive experiments across five benchmark datasets, demonstrating the effectiveness of our approach. Remarkably, CURE achieves an average mIoU of 39.6% with ViT-L/14, outperforming ClearCLIP by 4.3%. When integrated into baseline methods, it yields average mIoU gains of 2.8% and 7.7% with ViT-B/16 and ViT-L/14 backbones, respectively.

2 Related work

2.1 Vision-Language Models

Vision-Language Models (VLM) [1, 19, 24, 34, 52, 55] aim to bridge the gap between visual content and natural language by learning joint representations that align images with corresponding textual descriptions. Among these, CLIP [34] has become particularly influential. It is trained on 400 million image-text pairs using a contrastive learning objective and adopts a dual-encoder architecture with separate image and text encoders, projecting both modalities into a shared embedding space. This design enables strong zero-shot generalization across tasks such as image classification [27, 60], retrieval [30], and captioning [31, 38]. Despite these strengths, CLIP and similar models face challenges in dense prediction tasks such as semantic segmentation. Their image encoders are primarily optimized to capture global representations, typically through the use of a classification token (CLS) that aggregates high-level semantic information. While effective for image-level tasks, this global representation strategy often suppresses fine-grained, localized features that are essential for pixel-wise prediction. In this work, we seek to retain VLM's versatility while mitigating its global contextual bias, thereby enhancing the utility for dense segmentation.

2.2 Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation (OVSS) leverages pre-trained VLMs to segment arbitrary categories without task-specific annotations. Existing approaches can be grouped into training-based and training-free paradigms. Training-based methods typically employ a two-stage pipeline [26, 37, 51, 54], where a class-agnostic mask proposal network first generates candidate regions, followed by region-level classification via a pre-trained VLM [43]. Although these methods perform well, they are often computationally demanding and limited in their ability to model context. One-stage methods [11, 14, 47, 50] integrate segmentation and classification into a unified architecture, significantly improving efficiency. However, they still rely on annotated training data, limiting their generalization to unseen categories. In contrast, training-free OVSS methods directly exploit VLMs without task-specific fine-tuning, leveraging its ViT-based encoder to align image patch tokens with textual features. Nevertheless, recent work has shown that the final-layer representations of CLIP tend to capture excessive global context, which diminishes the granularity of local semantic information [16]. To mitigate this, several methods propose correlation calibration techniques. SCLIP [41] introduces a correlative self-attention module that averages query-to-query and key-to-key attention to enhance region-level correlation. ClearCLIP [22] modifies the final transformer block by removing residual connections and feedforward networks, an approach that has served as a foundational strategy for many subsequent methods. More recent studies [3, 21] further examine the impact of a globally biased context on segmentation performance. While sharing similar motivation, CURE adopts a distinct formulation. Instead of attention-level adjustments, we model the final representation as an entangled mixture of local semantics and global bias. By post-hoc decomposing this bias into explicit outliers and implicit context shifts, CURE effectively removes residual artifacts.

3 Method

3.1 Preliminary

The visual encoder in vision-language models (e.g., CLIP) typically adopts the Vision Transformer (ViT) architecture and consists of L Transformer blocks. The input image is first divided into non-overlapping patches, each projected into a patch token $x_i \in \mathbb{R}^d$, where d is the feature dimension. These tokens, together with a special classification (CLS) token x_{cls}, constitute the input token sequence $X = [x_{\text{cls}}, X_{\text{dense}}] = [x_{\text{cls}}, x_1, \ldots, x_N] \in \mathbb{R}^{(N+1)\times d}$. This sequence is then passed through a stack of residual attention blocks, allowing for the exchange of spatial and semantic information across tokens. The computation in the l-th Transformer block can be formally expressed as:

$$X^l_{sum} = X^{l-1} + \texttt{SA}\left(\texttt{LN}(X^{l-1})\right) \ , \tag{1}$$

$$X^l = X^l_{sum} + \texttt{FFN}\left(\texttt{LN}(X^l_{sum})\right) \ , \tag{2}$$

where LN denotes layer normalization, SA represents the self-attention layer, and FFN is the feed-forward network.

In the training-free OVSS setting, a straightforward baseline is to perform dense patch-level classification using the final-layer visual tokens. We leverage VLM's aligned vision-language embedding space, where class embeddings $X_{text} \in \mathbb{R}^{C\times d}$ are generated from prompt-enhanced textual labels, with C denoting the number of classes. The segmentation map $M \in \mathbb{R}^{N\times 1}$ is then produced by assigning each patch token to the class with the highest cosine similarity:

$$M = \arg\max_{C} \left(\cos(X^{L}_{dense}, X_{text})\right), \tag{3}$$

where X^{L}_{dense} represents the patch tokens from the final layer and $\cos(\cdot,\cdot)$ denotes cosine similarity.

3.2 Motivation

While VLMs achieve strong results in zero-shot classification [34], their visual encoders are less effective for dense prediction tasks such as semantic segmentation. This limitation arises from their pre-training objective, which emphasizes image-level representations via the CLS token. As a result, patch tokens in the final layers tend to capture excessive global context, compromising their ability to represent localized details [16]. Prior efforts [22,23] address this by modifying final-layer attention but fall short of fully resolving the issue.

To understand this phenomenon, we inspect attention maps and token statistics across layers. As shown in Figure 1, certain patch tokens exhibit attention patterns closely resembling that of the CLS token, suggesting residual global bias. In addition, these tokens also exhibit disproportionately high embedding norms, in line with prior observations [13, 46]. Quantitative analysis on ADE20K [58] and COCO-Stuff [7] in Figure 2 reveals that such tokens are common and tend to correspond to lower classification accuracy, indicating that global dominance is detrimental to segmentation quality.

These findings motivate our approach: *how can global contextual bias be suppressed and spatially localized semantics recovered, without modifying the pre-trained model or requiring additional training?*

3.3 Reject and Reconstruct

The preceding analysis highlights that globally biased tokens are a non-negligible factor limiting segmentation performance. While prior approaches, such as correlative self-attention (CSA) [22,41], attempt to correct attention patterns, they often retain these high-norm outlier tokens, i.e., local tokens continue to receive disproportionate attention and degrade spatial information.

To explicitly suppress this global bias, we first propose a simple yet effective module termed Reject and Reconstruct (RR). Unlike CSA-based mechanisms, RR directly removes globally biased tokens and restores local consistency through a two-stage process. In the *Reject* stage, tokens with embedding norms

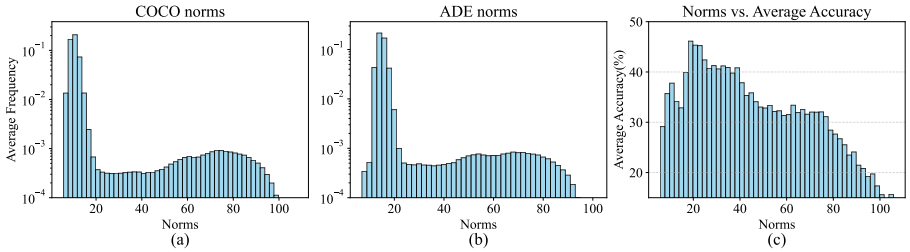

Fig. 2: Token-level analysis on COCO-Stuff [7] and ADE20K [58]. (a) Norm distribution of patch tokens on COCO-Stuff; (b) Norm distribution on ADE20K (tokens with a norm > 40 constitute 3.5% and 3.6% on COCO-Stuff and ADE20K, respectively); (c) Token-level classification accuracy as a function of token norm on ADE20K. In all cases, patch tokens are extracted from the penultimate block of the CLIP ViT-B/16.

above a predefined threshold are discarded. These high-norm tokens consistently exhibit global attention patterns and are treated as biased. In the *Reconstruct* stage, each rejected token is replaced by the mean of its spatial neighbors, leveraging the local continuity assumption in visual semantics to preserve structural coherence. Neighboring tokens are defined based on Chebyshev distances (with a threshold of 2). The reconstructed token at position $\mathbf{p} = (i, j)$ is computed as:

$$x_{\mathbf{p}}^{\text{rec}} = \frac{1}{|\mathcal{V}(\mathbf{p})|} \sum_{\mathbf{q} \in \mathcal{V}(\mathbf{p})} x_{\mathbf{q}}, \tag{4}$$

where $\mathcal{V}(\mathbf{p})$ denotes the set of positions of valid neighboring tokens determined by the Chebyshev distance criterion.

3.4 Global Contextual Debiasing

While the RR module suppresses explicitly biased tokens, it does not eliminate more subtle global artifacts that remain embedded in the representations of non-outlier tokens. These residual effects arise from the image-level contrastive pretraining objective of VLMs, which encourages patch tokens to encode shared global context. To further reduce this influence, we introduce a complementary module termed *Global Contextual Debiasing* (GCD).

GCD is built on the assumption that each patch token feature x_i can be decomposed into two components: a localized semantic part x_i^{sem} and a residual global artifact x_i^{bias} induced by the CLS token and global attention interactions. Explicitly separating these components is challenging due to the diversity of patch semantics across spatial locations. Drawing on principles from high-dimensional statistics [40], we assume that localized semantic features are approximately orthogonal because they correspond to statistically independent regions. In contrast, residual global artifacts are shared across tokens and therefore tend to align along a common direction in feature space. Based on this

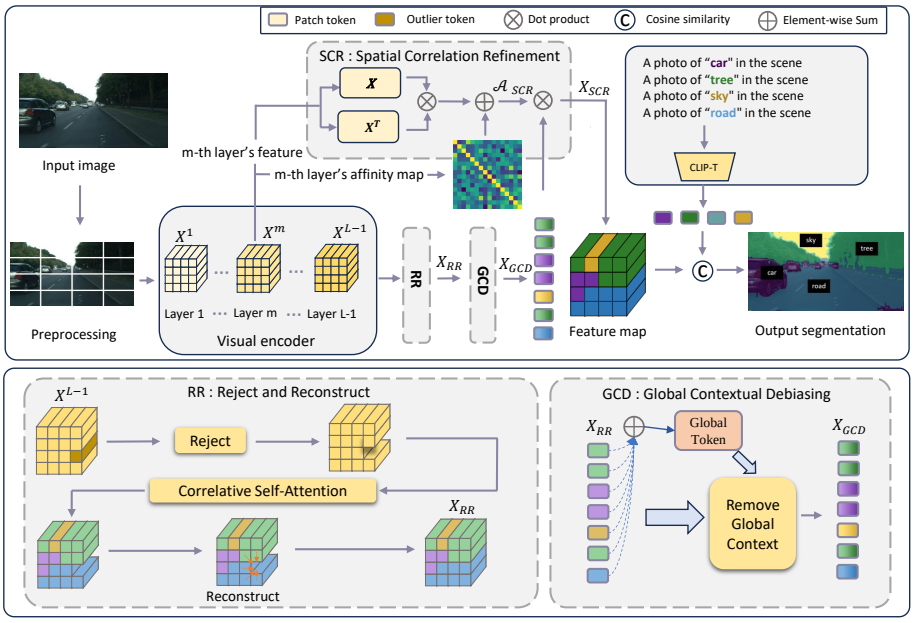

Fig. 3: Overview of the proposed CURE framework for addressing global contextual bias in VLMs. Visual patch tokens from the penultimate Transformer layer are first processed by the Reject and Reconstruct (RR) module, which discards high-norm outlier tokens exhibiting global attention artifacts and reconstructs them using neighboring tokens to preserve spatial semantics. The Global Contextual Debiasing (GCD) module then removes residual implicit bias from the remaining tokens by estimating and subtracting a global artifact component, obtained by averaging non-outlier embeddings. To further enhance spatial coherence, the Spatial Correlation Refinement (SCR) module computes an affinity matrix using intermediate-layer feature and value embeddings. This matrix captures spatial and semantic relationships and is used to refine the final-layer features. The resulting features are aligned with text embeddings via cosine similarity to produce the final segmentation map.

observation, we approximate the global bias vector by averaging the non-outlier patch features within an image. Specifically, we compute:

$$x_i = x_i^{sem} + x_i^{bias}, \quad \hat{x}_{i,RR}^{bias} \simeq \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{i,RR}, \\ x_{i,GCD} = x_{i,RR} - \alpha \cdot \hat{x}_{i,RR}^{\text{bias}}, \tag{5}$$

where $\tilde{x}_{i,RR}$ denotes the non-outlier token identified by the RR module, and $x_{i,GCD}$ represents the patch tokens after our GCD processing. α is a hyperparameter controlling the strength of the debiasing.

By estimating and subtracting the shared component, GCD reduces residual global influence while preserving localized and discriminative information. Together with RR, this two-stage framework addresses both explicit and implicit

sources of global contextual bias, yielding more robust patch representations for dense prediction tasks. Notably, a concurrent study [39] shows that modern vision models such as CLIP, SigLIP, and DINO exhibit partially linear factorization with low-rank, near-orthogonal per-concept components, which is consistent with our modeling assumption and provides empirical support for the proposed decomposition.

3.5 Spatial Correlation Refinement

As illustrated in Figure 1, outlier tokens emerge as early as the intermediate layers of VLM's visual encoder. Therefore, refining only the final-layer features is insufficient to fully mitigate the effects of global contextual bias. These early-stage outliers can distort spatial relationships among semantically related patches, adversely affecting segmentation quality. However, directly applying RR or GCD to intermediate layers risks disrupting VLM's embedding space, potentially impairing the zero-shot generalization.

To address this, we propose Spatial Correlation Refinement (SCR), which improves spatial coherence while preserving the pre-trained embedding structure. Rather than modifying intermediate features, SCR extracts spatial correlation patterns from a mid-level Transformer layer to guide the refinement of final-layer representations. Similar to prior work on spatial consistency and semantic disentanglement [25, 41], we compute two types of similarity: feature similarity, which encourages alignment among similar spatial regions, and value similarity, which helps separate distinct semantic areas. Formally, the operation is defined:

$$\begin{aligned} \mathcal{A}_{SCR} &= X^m \cdot (X^m)^T + V^m \cdot (V^m)^T, \\ X_{SCR} &= \mathcal{A}_{SCR} \cdot X_{GCD}, \end{aligned} \tag{6}$$

where X^m and V^m denote the patch features and value embeddings from the m-th intermediate Transformer layer. The affinity matrix $\mathcal{A}_{\text{SCR}} \in \mathbb{R}^{N \times N}$ encodes both spatial and semantic relationships across tokens. By combining these signals, SCR enhances intra-class consistency and inter-class discrimination, improving the quality and robustness of segmentation outputs.

4 Experiments

4.1 Experimental Settings

Datasets. We empirically evaluate the performance of CURE on five semantic segmentation benchmark datasets: COCO-Stuff(C-Stf) [7], PASCAL Context (PC59) [32], PASCAL VOC 2012 (VOC20) [15], Cityscapes [12], and ADE20K [58]. Further details regarding dataset statistics, data splits, and evaluation metrics are provided in the supplementary material.

Implementation Details. In line with previous work, we adopt CLIP [34] as the vision-language model and evaluate both the ViT-B/16 and ViT-L/14 architectures. During inference, input images are resized such that the shorter side is

448 pixels, and a sliding window of size 448×448 with a stride of 224 is applied, following the configuration in [22]. For Cityscapes dataset [12], a larger resize scale of 560 pixels is used to better preserve fine-grained details in high-resolution urban scenes. In this case, a smaller window size of 224×224 with a stride of 112 is employed, following standard practice. No post-processing is applied. The outlier threshold in the RR module is set to 40. For the GCD module, the debiasing strength parameter α is fixed at 0.25 for the ViT-B/16 backbone and 0.4 for the ViT-L/14 variant. Further detailed analyses are provided in the supplementary material. In the SCR module, mid-level features are extracted from the third Transformer block for ViT-B/16 and the sixth block for ViT-L/14, to account for differences in backbone depth. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Our pipeline is built on ClearCLIP [22], with CURE integrated into the inference process. For clarity, we omit the ClearCLIP prefix in the remainder of the text.

4.2 Experimental Results

We follow standard evaluation protocols and compare against both training-based methods (GroupViT [49], ReCLIP [42], TCL [8]) and training-free approaches (MaskCLIP [59], SCLIP [41], ClearCLIP [22], ResCLIP [53], LPOSS [36], SFP [21], SCCLIP [3] and FLOSS [4]). For fairness, all training-free baselines are reproduced and evaluated under identical settings. We also include vanilla CLIP to establish a baseline for zero-shot segmentation.

As shown in Table 1, the vanilla CLIP model with a ViT-B/16 backbone yields an average mIoU of 12.8%, indicating limited spatial precision. Training-free variants such as SCLIP [41] and ClearCLIP [22] improve upon this baseline by introducing correlative self-attention mechanisms, achieving mIoUs of 35.8% and 37.9%, respectively. However, these methods primarily focus on refining attention and do not explicitly address the residual global contextual bias involved in CLIP's image representations. In contrast, our proposed CURE framework incorporates a comprehensive three-stage strategy to suppress both explicit and implicit global artifacts. This leads to improved segmentation quality, with CURE achieving an average mIoU gain of 4.3% over SCLIP [41] and 2.2% over ClearCLIP [22].

A counter-intuitive observation is that OVSS performance using the ViT-L/14 backbone is generally lower than with ViT-B/16, despite its higher model capacity. This phenomenon is likely due to the fact that ViT-L/14's increased capacity leads to stronger encoding of global contextual bias. Since most existing methods rely on attention recalibration alone, they often fail to correct these highly biased representations. Consequently, the performance gap between ViT-L/14 and ViT-B/16 persists across prior methods. In contrast, CURE effectively suppresses the underlying bias and reduces this gap. On SCLIP and NACLIP, ViT-L/14 initially underperforms compared to ViT-B/16, but surpasses it after applying CURE. In particular, CURE boosts SCLIP on ViT-L/14 with an average mIoU gain of 15.7%, demonstrating CURE's ability to bridge backbone-level discrepancies driven by global bias.

Table 1: Quantitative comparison of open-vocabulary semantic segmentation performance (mIoU, %) across five datasets, evaluating our CURE against baseline approaches. Both ViT-B/16 and ViT-L/14 backbones are used to assess method robustness under varying model capacities. TF denotes training-free. Results for training-based baselines are cited from ReCLIP [42]. CURE consistently boosts existing training-free methods, including MaskCLIP [59], SCLIP [41], ClearCLIP [22], ResCLIP [53], and FLOSS [4], by effectively addressing both explicit and implicit global bias. CURE reduces the performance gap between ViT-B/16 and ViT-L/14, enabling ViT-L/14 to surpass ViT-B/16 in several cases, which prior methods often fail. "CLIP (Vanilla) w/ CURE" refers to applying our method to the original CLIP architecture, without modifying the final transformer block as other methods do.

Methods	Venue	TF	Arch.	VOC20	PC59	ADE20K	Citysc.	C-Stf	Avg.
GroupViT [49]	CVPR-22	×	ViT-B/16	79.7	23.4	9.2	11.1	15.3	27.7
TCL [8]	CVPR-23	×	ViT-B/16	77.5	30.3	14.9	23.1	19.6	33.1
ReCLIP [42]	CVPR-24	×	ViT-B/16	75.8	33.8	14.3	19.9	20.3	32.8
CLIP-DINOiser [45]	ECCV-24	×	ViT-B/16	80.8	36.0	20.1	31.1	24.6	39.2
CLIP-DINOiser w/ CURE				**83.8**	**38.6**	**21.5**	**34.2**	**26.4**	**41.5**
CLIP [34]	ICML-21	✓	ViT-B/16	41.7	9.1	2.2	6.6	4.4	12.8
CLIP (Vanilla) w/ CURE				**77.1**	**23.7**	**10.0**	**21.7**	**15.8**	**29.7**
GEM [6]	CVPR-24	✓	ViT-B/16	79.8	35.8	15.5	33.4	23.7	37.6
MaskCLIP [59]	ECCV-22	✓	ViT-B/16	60.2	26.1	12.2	25.7	16.4	28.1
MaskCLIP w/ CURE				**64.3**	**32.5**	**15.9**	**33.9**	**21.2**	**33.6**
SCLIP [41]	ECCV-24	✓	ViT-B/16	78.1	33.0	14.6	32.3	21.1	35.8
SCLIP w/ CURE				**79.2**	**38.2**	**18.6**	**37.0**	**25.5**	**39.7**
NACLIP [18]	WACV-25	✓	ViT-B/16	75.4	35.5	17.4	35.5	23.2	37.4
NACLIP w/ CURE				**76.9**	**37.9**	**18.7**	**38.0**	**24.9**	**39.3**
ResCLIP [53]	CVPR-25	✓	ViT-B/16	85.3	36.9	17.6	35.9	24.6	40.1
ResCLIP w/ CURE				**86.4**	**38.9**	**18.8**	**37.9**	**25.9**	**41.6**
FLOSS [4]	ICCV-25	✓	ViT-B/16	80.1	35.9	18.4	37.0	23.6	39.0
FLOSS w/ CURE				**82.6**	**37.7**	**19.6**	**39.6**	**25.1**	**40.9**
LPOSS [36]	CVPR-25	✓	ViT-B/16	78.8	37.8	21.8	37.3	25.9	40.3
LPOSS w/ CURE				**82.5**	**39.9**	**22.7**	**38.0**	**27.7**	**42.2**
SFP [21]	ICCV-25	✓	ViT-B/16	81.7	38.9	20.1	38.0	25.8	40.9
SFP w/ CURE				**82.6**	**39.9**	**20.7**	**39.6**	**26.4**	**41.8**
SCCLIP [3]	TIP-25	✓	ViT-B/16	84.3	40.1	20.1	41.0	26.6	42.4
SCCLIP w/ CURE				**84.7**	**40.3**	**20.4**	**42.3**	**26.6**	**42.9**
ClearCLIP [22]	ECCV-24	✓	ViT-B/16	80.9	35.9	16.7	32.0	23.9	37.9
CURE		✓	ViT-B/16	**82.1**	**38.7**	**18.4**	**35.7**	**25.8**	**40.1**
CLIP [34]	ICML-21	✓	ViT-L/14	15.7	4.5	1.3	3.2	2.4	5.4
MaskCLIP [59]	ECCV-22	✓	ViT-L/14	30.1	12.6	7.0	12.0	8.9	14.1
MaskCLIP w/ CURE				**57.2**	**29.3**	**17.2**	**26.3**	**19.5**	**29.9**
SCLIP [41]	ECCV-24	✓	ViT-L/14	60.3	20.5	7.0	21.1	13.1	24.4
SCLIP w/ CURE				**82.6**	**36.6**	**19.3**	**37.9**	**24.3**	**40.1**
NACLIP [18]	WACV-25	✓	ViT-L/14	76.8	31.6	16.8	31.4	20.8	35.5
NACLIP w/ CURE				**78.8**	**36.5**	**19.9**	**37.2**	**24.4**	**39.4**
ResCLIP [53]	CVPR-25	✓	ViT-L/14	86.6	32.9	16.8	33.7	22.3	38.5
ResCLIP w/ CURE				**87.4**	**36.1**	**18.7**	**38.6**	**24.6**	**41.1**
FLOSS [4]	ICCV-25	✓	ViT-L/14	77.6	32.8	17.5	31.3	21.4	36.1
FLOSS w/ CURE				**81.9**	**36.4**	**19.9**	**37.7**	**24.0**	**40.0**
SCCLIP [3]	TIP-25	✓	ViT-L/14	88.3	40.5	21.5	41.1	26.9	43.7
SCCLIP w/ CURE				**88.5**	**41.2**	**22.0**	**42.4**	**27.1**	**44.2**
ClearCLIP [22]	ECCV-24	✓	ViT-L/14	80.0	29.6	15.0	31.8	19.9	35.3
CURE		✓	ViT-L/14	**83.3**	**35.7**	**18.7**	**36.9**	**23.6**	**39.6**

Table 2: Ablation study of CURE using ViT-B/16, demonstrating the individual contributions of the RR, GCD, and SCR modules, as well as their cumulative effect.

RR	GCD	SCR	ADE20K	PC59	Citysc.	Avg.
-	-	-	16.7	35.9	32.0	28.2
✓	-	-	17.0	36.5	33.1	28.9
-	✓	-	17.3	36.3	33.8	29.1
-	-	✓	17.3	37.3	33.7	29.4
✓	✓	-	17.5	36.9	34.2	29.5
✓	-	✓	17.5	37.6	34.0	29.7
-	✓	✓	18.2	38.4	35.2	30.6
✓	✓	✓	**18.4**	**38.7**	**35.7**	**30.9**

4.3 Experimental Analysis

Ablation Study. To assess the individual contributions of each component in the proposed CURE framework, we perform a detailed ablation study using the ViT-B/16 backbone on three benchmark datasets: ADE20K, PASCAL Context, and Cityscapes. As shown in Table 2, each module of CURE provides a measurable improvement in segmentation performance. The results also highlight the complementary nature of these components: combining all three yields the highest mIoU across datasets, confirming the effectiveness.

Hyperparameter Analysis. CURE introduces three key hyperparameters: the outlier threshold in the RR module, the debiasing strength α in the GCD module, and the choice of the mid-level Transformer layer in the SCR module. To assess their effect on segmentation performance, we conduct experiments varying each parameter independently. The results are shown in Figure 4. For both the RR and GCD modules, performance remains stable across a range of threshold and α values, reflecting the robustness of the two-stage bias reduction mechanism. In contrast, the choice of SCR layer has a more substantial impact. Layers earlier than the third tend to yield suboptimal results due to insufficient semantic abstraction, while deeper layers may introduce accumulated global bias. Nevertheless, when appropriate mid-level layers are selected, the performance is stable, confirming the effectiveness of SCR.

Versatility across Different VLMs. In addition to CLIP, we evaluate the generalizability of CURE across other vision-language models. As shown in Table 4, CURE consistently improves the performance of the ClearCLIP baseline, regardless of the underlying pretraining strategy—BLIP, MetaCLIP, or OpenCLIP. Notably, it yields an average mIoU gain of +2.4% for both MetaCLIP and OpenCLIP, demonstrating its effectiveness as a general-purpose enhancement for diverse VLM backbones.

Effect in Addressing Explicit Global Bias Information. As illustrated in Figure 2, visual patch tokens tend to exhibit a negative correlation between embedding norm and token-level classification accuracy. Consistent with prior findings, high-norm tokens are often biased and correspond to outliers influenced

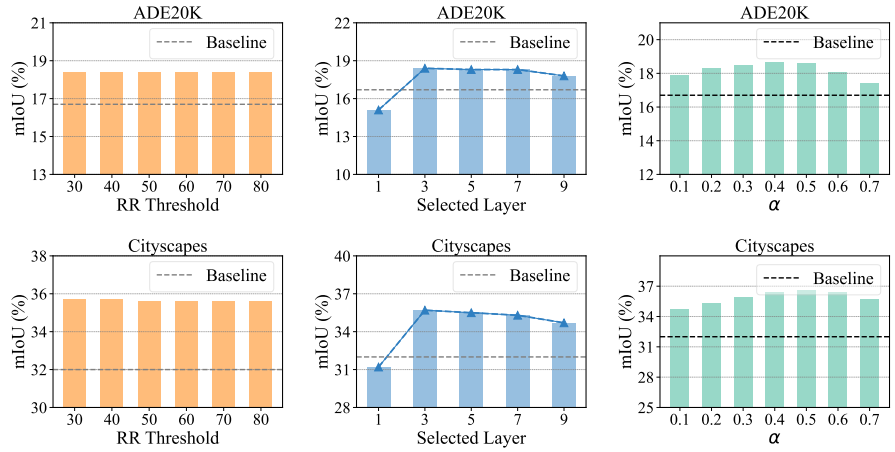

Fig. 4: Hyperparameter analysis of CURE. We assess the impact of the outlier threshold in the RR module, the debiasing strength α in the GCD module, and the selected Transformer layer in the SCR module. The results show that CURE remains robust across a broad range of values. However, selecting SCR layers that are too shallow or too deep can reduce performance due to limited semantic abstraction or increased global bias, respectively.

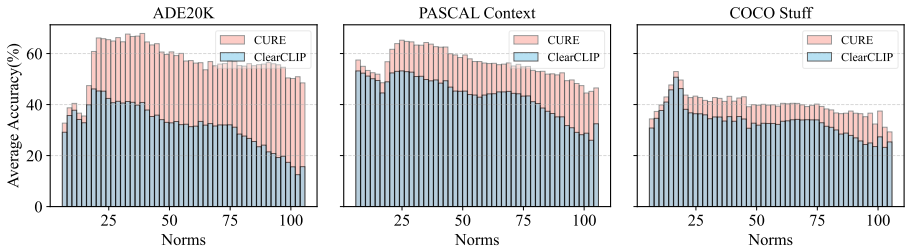

Fig. 5: Token-level classification accuracy comparison between ClearCLIP and CURE. High-norm tokens, often associated with global contextual bias, remain insufficiently corrected by prior attention calibration methods. CURE effectively mitigates this bias and enhances the local discriminative capacity of patch tokens.

by global image-level representations. To further evaluate the effectiveness of CURE in mitigating this bias, we analyze token-level classification accuracy on the ADE20K and PASCAL Context datasets, extending beyond standard mIoU metrics. As shown in Figure 5, CURE greatly improves the accuracy of outlier tokens, offering empirical evidence of its ability to restore locally discriminative features. More results are included in the appendix.

Mitigating Bias in Background Regions. A persistent challenge in training-free OVSS is the model's bias toward frequent foreground objects, often resulting in poor recognition of globally dominant background regions. This imbalance de-

Table 3: Evaluation (mIoU, %) on semantic segmentation tasks including the background class.

Method	Backbone	VOC21	PC60	Object	Avg.
ClearCLIP	ViT-B/16	51.8	32.6	33.0	39.1
+CURE	ViT-B/16	**56.1**	**35.2**	**36.1**	**42.5**
ResCLIP	ViT-B/16	61.6	33.0	35.2	43.3
+CURE	ViT-B/16	**65.1**	**34.9**	**36.4**	**45.5**
ClearCLIP	ViT-L/14	46.1	26.7	30.1	34.3
+CURE	ViT-L/14	**52.6**	**31.0**	**34.8**	**39.5**
ResCLIP	ViT-L/14	54.4	29.2	31.9	38.5
+CURE	ViT-L/14	**59.7**	**32.2**	**33.7**	**41.9**

Table 4: Evaluation (mIoU, %) on semantic segmentation tasks with different VLM backbones.

Pre-training	ADE20K	COCO	PC59	Avg.
BLIP + ClearCLIP	13.1	21.2	31.0	21.8
BLIP + CURE	**13.9**	**22.6**	**34.2**	**23.6**
MetaCLIP + ClearCLIP	17.4	23.5	34.8	25.2
MetaCLIP + CURE	**18.9**	**25.9**	**38.0**	**27.6**
OpenCLIP + ClearCLIP	18.9	23.1	34.1	25.4
OpenCLIP + CURE	**20.7**	**25.4**	**37.3**	**27.8**

Table 5: Evaluation (mIoU, %) on long-tail categories.

Method	City	City (tail)	ADE	ADE(tail)	Avg.
ClearCLIP	32.0	26.4	16.7	14.6	22.4
+CURE	**35.7**	**27.1**	**18.4**	**16.0**	**24.3**
SCLIP	32.3	25.3	14.6	12.9	21.3
+CURE	**37.0**	**27.2**	**18.6**	**16.1**	**24.7**

Table 6: Effectiveness of GCD vs. direct [CLS] token subtraction.

Method	ADE20K	PC59	Citysc.	Avg.
CURE (w/o GCD)	17.5	37.6	34.0	29.7
Direct [CLS] Sub.	17.9	36.9	34.7	29.8
CURE (Full)	**18.4**	**38.7**	**35.7**	**30.9**

grades segmentation performance in scenes rich in contextual content. As shown in Table 3, CURE consistently enhances performance on benchmarks that explicitly include a background class, surpassing all baselines. These improvements demonstrate CURE's effectiveness in suppressing global contextual bias and increasing segmentation accuracy in background-dominated areas.

Robustness to Long-tail Categories. Long-tail categories typically occupy small spatial regions and exhibit high intra-class variation, making them difficult to segment without strong localized representations. This challenge is further amplified in training-free OVSS, as these data are rarely represented in the VLM's pretraining data. Following the class split from [44] for Cityscapes, and identifying the 130 ADE20K classes with less than 1% pixel proportion as long-tail, we evaluate CURE's performance on these difficult cases. As shown in Table 5, CURE consistently upgrades prior methods, with an average of 1.9% and 3.4% mIoU improvement to ClearCLIP and SCLIP, highlighting its effectiveness in recovering fine-grained, discriminative features for long-tail classes.

GCD vs. vanilla `[CLS]` Subtraction. A straightforward alternative to GCD is to subtract the CLS token. However, as illustrated in Table 6, this vanilla strategy leads to unstable results, e.g., a 0.7% mIoU drop on PC59. The reason lies in the role of the CLS token during pretraining. It is optimized for image-level contrastive alignment and therefore encodes dominant, task-relevant semantics of the entire image rather than a uniform global shift shared across spatial locations. Consequently, it is an unreliable proxy for the residual artifacts embedded in patch tokens, and subtracting it may remove useful semantic information. In contrast, GCD estimates the shared bias by averaging non-outlier patch features, forming an empirical centroid. Since this estimate is derived directly from spa-

Table 7: Efficiency analysis of CURE and its subcomponents. Computational cost (GFLOPs) and inference speed (IPS: Images Per Second) are measured using a ViT-B/16 encoder with 224×224 input resolution on a single NVIDIA RTX 3090 GPU.

Metric	ClearCLIP	+RR	+GCD	+SCR	CURE
FLOPs (G) ↓	34.9	34.9	34.9	35.0	35.0
Speed (IPS)↑	35.3	32.7	33.8	34.8	32.3

tial tokens, it better reflects the common contextual component across patches, leading to more stable and consistent improvements across benchmarks.

Efficiency Analysis. The computational overhead introduced by CURE is minimal, as all its components (RR, GCD, and SCR) operate exclusively on the final-layer features of the ViT encoder. GCD involves a lightweight global averaging step, RR performs reconstruction via a single parallel convolution, and SCR adds an efficient affinity-based aggregation. We evaluate the computational cost on a single NVIDIA RTX 3090 GPU using input images of resolution 224×224. As shown in Table 7, the complete CURE framework incurs negligible additional overhead, demonstrating its suitability for efficient training-free deployment.

Visualization. In Figure 6, we present qualitative comparisons between CURE and several baseline methods. CURE produces segmentation maps that are generally cleaner and more spatially consistent. Object boundaries are better preserved, and large homogeneous regions contain fewer spurious activations. In contrast, baseline methods often exhibit fragmented predictions or excessive background responses, which are consistent with insufficient suppression of global contextual bias. At the same time, these visual results also reveal certain limitations. In scenes where foreground objects occupy a large portion of the image or share similar appearance with the background, CURE may still produce partial under-segmentation or slight boundary smoothing. This behavior is related to the global subtraction mechanism in GCD, which could attenuate features that are strongly correlated across the image. In addition, although RR reduces extreme activations, small and fine-grained structures may remain challenging when their representations are weak relative to dominant regions. These observations indicate that, while CURE improves robustness to contextual bias, handling highly dominant foregrounds and subtle local details remains an open direction for further refinement.

5 Conclusion

In this work, we study open-vocabulary semantic segmentation using pre-trained vision–language models. Unlike prior approaches that mainly rely on attention map calibration to reduce biased signals, we explicitly target global contextual bias and aim to recover spatially localized semantics. To this end, we propose CURE (Contextual Debiasing and Unbiased Refinement), a framework that addresses both explicit and implicit forms of global bias. CURE consists of three

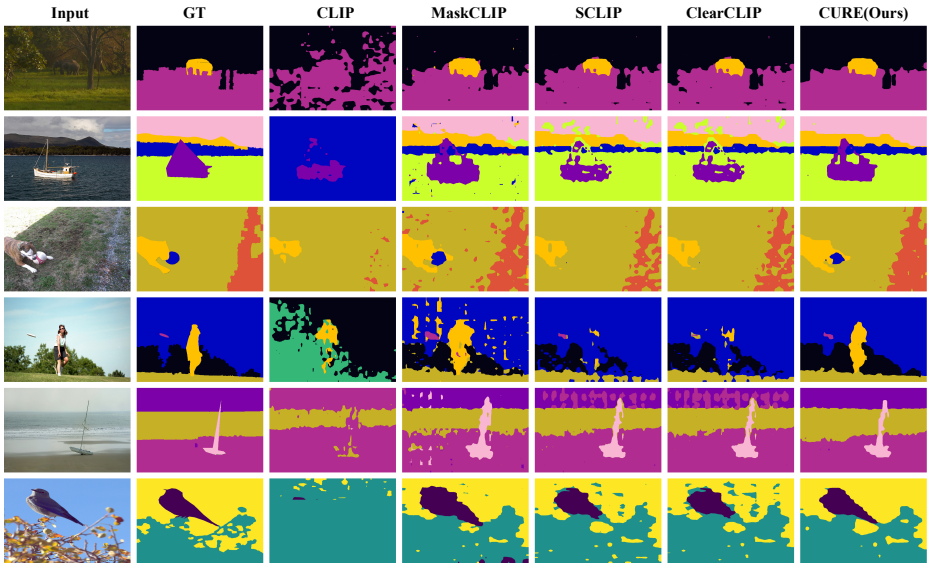

Fig. 6: Qualitative comparison of segmentation results across different methods.

components: outlier token rejection and reconstruction, global contextual debiasing, and spatial correlation refinement. Extensive experiments on five benchmarks demonstrate that CURE consistently improves segmentation accuracy and robustness over existing training-free methods. On average, it achieves mIoU gains of 2.8% and 7.7% with ViT-B/16 and ViT-L/14 backbones, respectively. Notably, CURE narrows the performance gap between the two backbones and enables ViT-L/14 to surpass ViT-B/16 in several settings. Since prior methods often fail to unlock the potential of larger backbones due to stronger global bias, this result highlights the effectiveness of CURE in suppressing contextual bias and restoring locally discriminative representations.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2024YFE0211000), in part by the National Natural Science Foundation of China (No. 62372329, 62506263, 62506264), in part by the Shanghai Scientific Innovation Foundation (No. 23DZ1203400), in part by the China Postdoctoral Science Foundation (No. BX20250383, GZB20250385, 2025M771530, 2025M771539), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

References

1. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. Advances in neural information processing systems **33**, 25–37 (2020)
2. Badrinarayanan, V., Kendall, A., SegNet, R.C.: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv **1511**(5) (2022)
3. Bai, S., Liu, Y., Han, Y., Zhang, H., Tang, Y., Zhou, J., Lu, J.: Self-calibrated clip for training-free open-vocabulary segmentation. IEEE Transactions on Image Processing (2025)
4. Benigmim, Y., Fahes, M., Vu, T.H., Bursuc, A., de Charette, R.: Floss: Free lunch in open-vocabulary semantic segmentation. arXiv preprint arXiv:2504.10487 (2025)
5. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)
6. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3828–3837 (2024)
7. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
8. Cha, J., Mun, J., Roh, B.: 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. pp. 11165–11174 (2023)
9. Chen, J., Zhu, D., Qian, G., Ghanem, B., Yan, Z., Zhu, C., Xiao, F., Culatana, S.C., Elhoseiny, M.: Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 699–710 (2023)
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
11. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4113–4123 (2024)
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
13. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
14. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)
15. Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep **8**(5), 2–5 (2011)

16. Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., Wilson, A.G., Goldstein, T.: What do vision transformers learn? a visual exploration. arXiv preprint arXiv:2212.06727 (2022)
17. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
18. Hajimiri, S., Ben Ayed, I., Dolz, J.: Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 5061–5071 (February 2025)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
20. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware clip representations for zero-shot segmentation. Advances in Neural Information Processing Systems **36**, 35631–35653 (2023)
21. Jin, S., Yu, S., Zhang, B., Sun, M., Dong, Y., Xiao, J.: Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20291–20300 (2025)
22. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: ECCV. pp. 143–160. Springer (2024)
23. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In: European Conference on Computer Vision. pp. 70–88. Springer (2024)
24. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
25. Li, Y., Wang, H., Duan, Y., Zhang, J., Li, X.: A closer look at the explainability of contrastive language-image pre-training. Pattern Recognition p. 111409 (2025)
26. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023)
27. Lin, Y., Chen, M., Zhang, K., Li, H., Li, M., Yang, Z., Lv, D., Lin, B., Liu, H., Cai, D.: Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3513–3521 (2024)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
29. Luo, J., Khandelwal, S., Sigal, L., Li, B.: Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4029–4040 (2024)
30. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In: Proceedings of the 30th ACM international conference on multimedia. pp. 638–647 (2022)
31. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)

32. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: 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. pp. 891–898 (2014)
33. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1520–1528 (2015)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
35. Shin, H., Kim, C., Hong, S., Cho, S., Arnab, A., Seo, P.H., Kim, S.: Towards open-vocabulary semantic segmentation without semantic labels. Advances in Neural Information Processing Systems **37**, 9153–9177 (2024)
36. Stojnić, V., Kalantidis, Y., Matas, J., Tolias, G.: Lposs: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9794–9803 (2025)
37. Sun, S., Li, R., Torr, P., Gu, X., Li, S.: Clip as rnn: Segment countless visual concepts without training endeavor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13171–13182 (2024)
38. Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D., Li, X.: Clip4caption: Clip for video caption. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4858–4862 (2021)
39. Uselis, A., Dittadi, A., Oh, S.J.: Compositional generalization requires linear, orthogonal representations in vision embedding models (2026), `https://arxiv.org/abs/2602.24264`
40. Vershynin, R.: High-dimensional probability: An introduction with applications in data science, vol. 47. Cambridge university press (2018)
41. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. In: European Conference on Computer Vision. pp. 315–332. Springer (2024)
42. Wang, J., Kang, G.: Learn to rectify the bias of clip for unsupervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4102–4112 (2024)
43. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. Advances in Neural information processing systems **33**, 17721–17732 (2020)
44. Wang, Y., Fei, J., Wang, H., Li, W., Bao, T., Wu, L., Zhao, R., Shen, Y.: Balancing logit variation for long-tailed semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19561–19573 (2023)
45. Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., Pérez, P.: Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In: European Conference on Computer Vision. pp. 320–337. Springer (2024)
46. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023)
47. Xie, B., Cao, J., Xie, J., Khan, F.S., Pang, Y.: Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3426–3436 (2024)

48. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
49. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18134–18144 (2022)
50. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2945–2954 (2023)
51. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
52. Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19163–19173 (2022)
53. Yang, Y., Deng, J., Li, W., Duan, L.: Resclip: Residual attention for training-free dense vision-language inference. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29968–29978 (2025)
54. Yilmaz, G., Peng, S., Engelmann, F., Pollefeys, M., Blum, H.: Opendas: Domain adaptation for open-vocabulary segmentation. arXiv e-prints pp. arXiv–2405 (2024)
55. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
56. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
57. Zhang, F., Zhou, T., Li, B., He, H., Ma, C., Zhang, T., Yao, J., Zhang, Y., Wang, Y.: Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. Advances in Neural Information Processing Systems **36**, 73652–73665 (2023)
58. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
59. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
60. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
61. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11175–11185 (2023)
62. Zhu, C., Chen, L.: A survey on open-vocabulary detection and segmentation: Past, present, and future. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
63. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in neural information processing systems **36**, 19769–19782 (2023)