LoGoSeg: Integrating Local and Global Features for Open-Vocabulary Semantic Segmentation

Junyang Chen[1,2], **Xiangbo Lv**[1,2,3], **Zhiqiang Kou**[1,2], **Xingdong Sheng**[3], **Ning Xu**[1,2], **Yiguo Qiao**[1,2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[2]Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 211189, China
[3]Lenovo Research, Shanghai 201203, China
{chenjunyang, zhiqiang_kou, xning, yqiao}@seu.edu.cn, {lvxb6, shengxd1}@lenovo.com

Abstract

Open-vocabulary semantic segmentation (OVSS) extends traditional closed-set segmentation by enabling pixel-wise annotation for both seen and unseen categories using arbitrary textual descriptions. While existing methods leverage vision-language models (VLMs) like CLIP, their reliance on image-level pretraining often results in imprecise spatial alignment, leading to mismatched segmentations in ambiguous or cluttered scenes. However, most existing approaches lack strong object priors and region-level constraints, which can lead to object hallucination or missed detections, further degrading performance. To address these challenges, we propose LoGoSeg, an efficient single-stage framework that integrates three key innovations: (i) an object existence prior that dynamically weights relevant categories through global image-text similarity, effectively reducing hallucinations; (ii) a region-aware alignment module that establishes precise region-level visual-textual correspondences; and (iii) a dual-stream fusion mechanism that optimally combines local structural information with global semantic context. Unlike prior works, LoGoSeg eliminates the need for external mask proposals, additional backbones, or extra datasets, ensuring efficiency. Extensive experiments on six benchmarks (A-847, PC-459, A-150, PC-59, PAS-20, and PAS-20^b) demonstrate its competitive performance and strong generalization in open-vocabulary settings.

1 Introduction

Semantic segmentation, a fundamental computer vision task that assigns semantic labels to each image pixel, achieves high accuracy on fixed-category benchmarks (Cheng et al. 2022; Chen et al. 2018; Srivastava and Sharma 2024). However, these conventional methods suffer from overfitting, are constrained by predefined label sets, and fail to generalize to unseen classes. To address these limitations, open-vocabulary semantic segmentation (OVSS) has recently emerged, enabling pixel-level prediction for arbitrary categories specified via textual prompts (Zhao et al. 2017; Xian et al. 2019; Lüddecke and Ecker 2022). This paradigm significantly broadens the applicability of semantic segmentation beyond closed-set constraints.

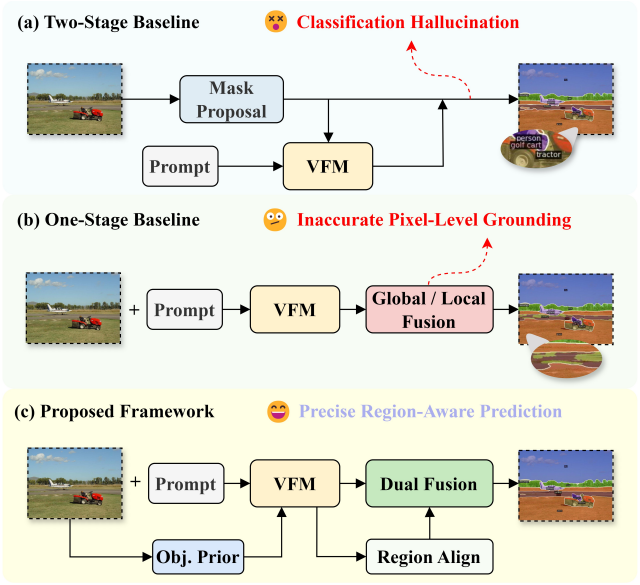

Figure 1: Comparison of open-vocabulary segmentation frameworks. (Top) Two-stage methods rely on external mask proposals, often causing hallucinations. (Middle) One-stage methods are more efficient but struggle with pixel-level grounding. (Bottom) LoGoSeg integrates object priors, region alignment, and dual fusion to improve cross-modal consistency and segmentation quality.

The core challenge of OVSS lies in its requirement to handle open-domain categories during inference—a capability fundamentally absent in traditional vision-only backbone networks. To address this, current research predominantly employs vision-language models (VLMs) such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), EVA-CLIP (Sun et al. 2023) and the SigLIP series (Tschannen et al. 2025; Zhai et al. 2023), leveraging their zero-shot recognition capabilities acquired through large-scale image-text pretraining (Sun et al. 2025; Yang et al. 2025). However, these models' image-level supervision paradigm creates inherent difficulties in adapting to pixel-level segmentation tasks (Zhou, Loy, and Dai 2022; Zhou et al.

*Corresponding author.

2019). Specifically, the pretraining process lacks explicit constraints for aligning textual concepts with visual regions, resulting in blurred segmentation boundaries and uncertain category predictions—limitations that become particularly pronounced when dealing with unseen or rare categories in open-vocabulary scenarios. Furthermore, the absence of object priors and region-aware constraints may lead to false positives or missed detections in complex scenes, significantly compromising segmentation accuracy.

To address these challenges, researchers have developed two distinct methodological frameworks: two-stage (Liang et al. 2023; Ding, Wang, and Tu 2022; Xu et al. 2023c; Ding et al. 2022; Xu et al. 2022b; Ghiasi et al. 2022; Xu et al. 2023a; Yu et al. 2023; Jiao et al. 2024; Wang et al. 2025) and single-stage (Dong et al. 2023; Cho et al. 2024; Zhou et al. 2023; Xu et al. 2023b; Xie et al. 2024; Shan et al. 2024; Sun et al. 2025) approaches. The two-stage paradigm, exemplified by mask proposal generators such as SAM (Kirillov et al. 2023; Ravi et al. 2024; Zhang, Liu, and Tang 2025) and MaskFormer (Cheng et al. 2022), predict category-agnostic mask proposals and then classify them using the zero-shot capability of frozen CLIP models. While achieving competitive accuracy, the efficiency is hindered by separate mask generation and classification stages. As shown in Fig. 1(a), they fail to fully exploit contextual information and depend heavily on mask quality, which limits their generalization.

Single-stage approaches, by contrast, directly adapt VLMs for pixel-level segmentation, simultaneously enhancing efficiency while leveraging the models' inherent segmentation capabilities. While recent advances like SAN (Xu et al. 2023b), SED (Xie et al. 2024), and CAT-Seg (Cho et al. 2024) have demonstrated promising results, Fig. 1(b) reveals critical limitations: an imbalance between local and global feature integration, insufficient region-level alignment, and unaddressed classification hallucinations.

To overcome the limitations of current methods, we propose LoGoSeg, a unified single-stage framework that addresses three core challenges in OVSS: classification hallucinations, misaligned pixel-level supervision, and insufficient fusion of local and global cues. The proposed architecture incorporates three novel components: first, an adaptive object existence prior computed from global image-text similarity dynamically adjusts category weights to suppress false positives; second, our region-aware alignment module establishes precise correspondences between visual regions and textual concepts through localized similarity computation; third, a dual-stream fusion mechanism jointly encodes fine-grained spatial structures and high-level semantic context in a unified representation space. Our contributions are summarized below:

(i) Region-aware alignment with object prior: Our strategy significantly reduces category hallucinations while improving pixel-level text-visual alignment through explicit region-level constraints.

(ii) Dual-stream feature fusion: The proposed framework effectively combines detailed local spatial context with comprehensive global semantics, overcoming the limitations of previous single-stream approaches.

(iii) Unified single-stage framework: LoGoSeg integrates these innovations into a single-stage pipeline that demonstrates superior performance across six standard benchmarks, achieving state-of-the-art results in both accuracy and computational efficiency.

2 Method

We propose LoGoSeg, a region-aware framework for OVSS, as illustrated in Fig. 2. Given an input image and a set of category prompts, the model first extracts image-level visual and textual features using CLIP. The prior-guided regional alignment module (**Section 2.2**) then measures similarity between regional visual features and adjusted text embeddings to provide category-aware guidance. These aligned features are subsequently passed to the cross-modal fusion module (**Section 2.3**), which integrates local and global cues via parallel architectures capturing directional spatial details and long-range contextual dependencies. Finally, the transformer-based decoder (**Section 2.4**) aggregates the fused features and predicts the final segmentation masks using learnable queries. Comprehensive details of each module are provided in the following subsections.

2.1 Preliminary

Open-vocabulary semantic segmentation (OVSS) performs dense prediction of text-defined categories at the pixel level. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a set of N text prompts $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$, the task outputs a segmentation map $\mathbf{Y} \in \{1, \ldots, N\}^{H\times W}$, where each pixel is assigned the index of its corresponding prompt. We use CLIP as the backbone, with image encoder $f_{\text{img}}(\cdot)$ producing $\mathbf{V} = f_{\text{img}}(\mathbf{I}) \in \mathbb{R}^{C\times H\times W}$ and text encoder $f_{\text{text}}(\cdot)$ producing prompt embeddings $\mathbf{T} = f_{\text{text}}(\mathcal{T}) \in \mathbb{R}^{N\times P\times C}$. However, as CLIP is trained only at the image level, these embeddings lack direct pixel correspondence.

Notation. Let B denote batch size, C feature channels, $H \times W$ spatial dimensions, N the number of categories, P prompts per category, K number of regions, D guidance dimension, T number of queries, and L_t pooled text tokens. The text embedding tensor is denoted as boldface $\mathbf{T}$, while the prompt set is calligraphic $\mathcal{T}$. We use $\|\cdot\|$ for ℓ_2 norm, $[a \parallel b]$ for channel-wise concatenation, and distinguish softmax_k (normalization over regions) from softmax_n (normalization over categories), respectively.

2.2 Prior-Guided Regional Alignment

To improve category discriminability and reduce semantic confusion under open-vocabulary settings, we propose a prior-guided regional alignment module that integrates object priors with region-level vision-language correlation.

Object Prior Estimation. Let $\mathbf{V} \in \mathbb{R}^{C\times H\times W}$ denote the CLIP-extracted image feature map, and $\mathbf{T} \in \mathbb{R}^{N\times P\times C}$ represent the prompt-based semantic text embeddings for N categories with P diverse prompts each. We compute the average prompt representation for each category as $\bar{\mathbf{T}}_n = \frac{1}{P}\sum_{p=1}^{P} \mathbf{T}_{n,p}$. The object-level prior is estimated by measuring the similarity between image features and the corre-

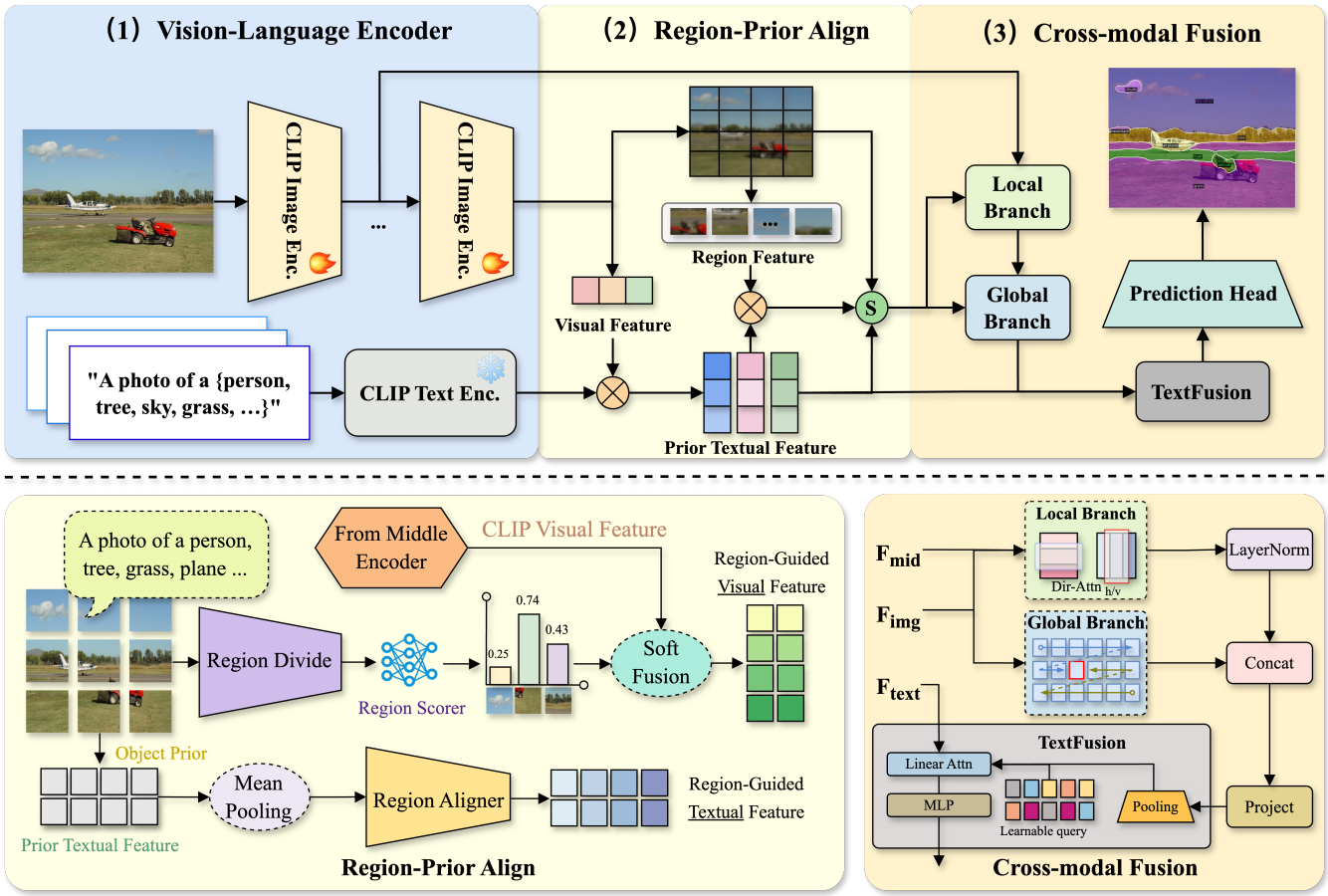

Figure 2: Overview of LoGoSeg. CLIP encoders extract multi-level visual and textual embeddings. A prior-guided alignment module uses a lightweight MLP to score regions for category-aware guidance and a Region Aligner to tightly align visual and textual features. A dual-branch fusion integrates local directional self-attention with global state-space modeling. A transformer with learnable queries then performs fine-grained cross-modal fusion via linear attention, and a hierarchical decoder with learnable upsampling and guidance-modulated convolutions yields the final segmentation map.

sponding text embedding:

$$p_n^{\text{prior}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\langle \mathbf{V}_{:,i,j}, \bar{\mathbf{T}}_n \rangle}{\|\mathbf{V}_{:,i,j}\| \cdot \|\bar{\mathbf{T}}_n\|}. \tag{1}$$

This score quantifies the likelihood of category n, serving as a λ-scaled semantic prior. The weighted prompt center is then derived by reweighting the embeddings:

$$\hat{\bar{\mathbf{T}}}_n = \frac{1}{P} \sum_{p=1}^{P} p_n^{\text{prior}} \, \mathbf{T}_{n,p}. \tag{2}$$

Region-Level Textual Guidance. The feature map is partitioned into K non-overlapping regions $\{R_k\}$, where $\mathbf{v}_k$ is mean-pooled within R_k. Given the λ-weighted prompt center $\hat{\bar{\mathbf{T}}}_n$, the region-category similarity matrix is defined as:

$$A_{k,n} = \frac{\langle \mathbf{v}_k, \hat{\bar{\mathbf{T}}}_n \rangle}{\|\mathbf{v}_k\| \cdot \|\hat{\bar{\mathbf{T}}}_n\|}, \quad A \in \mathbb{R}^{K \times N}. \tag{3}$$

This similarity matrix is transformed into region-aware weights via temperature-scaled softmax normalization:

$$w_{k,n} = \text{softmax}_k\big(\tau \, A_{k,n}\big), \quad \sum_{k=1}^{K} w_{k,n} = 1, \tag{4}$$

where τ adjusts the sharpness and $w_{k,n}$ quantifies the contribution of region k to category n. The visual prototype for category n is then obtained as:

$$\mathbf{m}_n = \sum_{k=1}^{K} w_{k,n} \, \mathbf{v}_k, \quad \mathbf{m}_n \in \mathbb{R}^C, \tag{5}$$

where $\mathbf{m}_n$ represents the aggregated visual feature for category n. Finally, we form the region-level textual guidance by projecting the concatenation of the prototype and the weighted text center:

$$g_{\text{region}}[n] = W_t \left[\mathbf{m}_n \parallel \hat{\bar{\mathbf{T}}}_n\right] + \mathbf{b}_t, \quad g_{\text{region}}[n] \in \mathbb{R}^D. \tag{6}$$

Region-Level Visual Guidance. For each region R_k, a trainable two-layer MLP (Linear-GELU-Linear) generates spatial attention logits, which are normalized via softmax to produce aggregation weights. These weights pool visual tokens into region embeddings $g_{\text{image}}^{(k)}$. An optional linear projection then aligns its dimension to that of g_{region} for seamless cross-modal fusion.

Region-aware Guidance Integration. The fusion process combines visual and textual cues through:

$$\begin{cases} G^{(k)} = \alpha_k \, g_{\text{image}}^{(k)} + (1-\alpha_k)\, \tilde{g}_{\text{text}}^{(k)}, \\ \alpha_k = \dfrac{\exp\big(\|g_{\text{image}}^{(k)}\|\big)}{\exp\big(\|g_{\text{image}}^{(k)}\|\big) + \exp\big(\|\tilde{g}_{\text{text}}^{(k)}\|\big)}, \\ \tilde{g}_{\text{text}}^{(k)} = W_g\Big(\sum_{n=1}^{N} \text{softmax}_n(\beta\, A_{k,n})\, g_{\text{region}}[n]\Big), \\ G_{:,i,j} = G^{(k)}, \quad \forall (i,j) \in R_k, \end{cases} \tag{7}$$

where $G \in \mathbb{R}^{B\times H\times W\times D}$ is the region-aware guidance tensor and β controls the category-wise softmax sharpness.

As illustrated in Fig. 2, the proposed module incorporates both global semantic priors and localized region-level alignment. By combining object-aware textual refinement and region-scored visual fusion, the model reduces hallucination and enhances fine-grained category discrimination, especially in cluttered or visually ambiguous scenes.

2.3 Contextual Cross-modal Fusion

To overcome the limitations of global feature representations in spatial precision and fine-grained alignment, we propose a contextual cross-modal fusion module unifying spatial detail with semantic abstraction, which is essential for accurate open-vocabulary segmentation. It integrates local structure and global semantics through three key components: first, a direction-aware attention mechanism that captures structured spatial information; second, a state-space modeling branch for long-range global context; and third, an adaptive fusion unit guided by both appearance and language cues.

Given region-aware guidance G and CLIP middle-layer visual features $F_{\text{mid}} \in \mathbb{R}^{B\times H\times W\times C}$, we decompose both tensors evenly along channel dimensions into horizontal/vertical halves: $G = (G_h, G_v)$, $F_{\text{mid}} = (F_h, F_v)$. We concatenate each pair $(F_h \parallel G_h)$ and $(F_v \parallel G_v)$ and apply rectangular self-attention—MHSA_h on the horizontal branch, MHSA_v on the vertical to capture directional dependencies. The two branch outputs are then concatenated:

$$X_{\text{local}} = \big[\text{MHSA}_h(F_h \parallel G_h),\, \text{MHSA}_v(F_v \parallel G_v)\big]. \tag{8}$$

The local and global branches are then fused as:

$$X_{\text{fused}} = \mathcal{F}_{\text{proj}}\big([\, X_{\text{local}} \parallel \text{SS2D}(G)\,]\big), \tag{9}$$

where $\text{SS2D}(\cdot)$ denotes the 2D state-space global-context module (Liu et al. 2024), $\mathcal{F}_{\text{proj}}$ is a 1×1 convolution for channel alignment.

This fusion strategy enables simultaneous encoding of fine-grained spatial structures and long-range semantic relationships, significantly enhancing cross-modal representation quality. By adaptively combining local and global cues, the model achieves superior pixel-text alignment and region-level discrimination, particularly in complex scenes.

Channel adaptation employs a sequential structure consisting of a 1×1 convolution, a depth-wise convolution, a gated linear unit block, and a second 1×1 projection, each followed by a residual connection and layer normalization. For text integration, we introduce T learnable query embeddings $Q_{\text{lrn}} \in \mathbb{R}^{B\times T\times D}$ and pooled text embeddings $E_{\text{text}} \in \mathbb{R}^{B\times L_t\times D}$, where text features are linearly projected to dimension D. We then form the key and value matrices as follows:

$$K = \big[X_{\text{fused}} \parallel E_{\text{text}}\big], \quad V = \big[X_{\text{fused}} \parallel E_{\text{text}}\big], \tag{10}$$

the linear attention $\text{Attn}(Q_{\text{lrn}}, K, V)$ is adopted, followed by token-wise gated fusion and a residual MLP.

In summary, our fusion module hierarchically integrates local patterns, global context, and language semantics. This approach enables precise cross-modal alignment and improves region-level discrimination in complex scenes.

2.4 Decoder and Loss Function

Decoder. We construct a correlation tensor $F \in \mathbb{R}^{B\times C\times T\times H\times W}$ by interacting T learnable queries with the fused feature map. The decoding procedure is as follows:

$$F_{i+1} = \text{Decoder}_i\big(\text{Upsample}(F_i),\, S_i\big), \tag{11}$$

where for $i{=}1, 2$, S_i denotes multi-scale *spatial* guidance features extracted from intermediate CLIP encoder activations, resized to match the decoder's resolution. Each decoding stage performs learnable upsampling followed by a guidance-modulated convolution, enabling precise reconstruction of fine-grained structures.

Finally, the decoder output is projected to per-class logits and then upsampled to the original input resolution (H_0, W_0) via bilinear interpolation:

$$\hat{Y}_{\text{final}} = \text{Interp}(\hat{Y}[:, :H, :W], \text{size} = (H_0, W_0)). \tag{12}$$

Loss Function. A pixel-wise multi-label binary cross-entropy loss is employed,

$$\begin{aligned} \mathcal{L}_{\text{BCE}} = -\frac{1}{\sum_{i,j} M_{i,j}} \sum_{i,j,n} M_{i,j} \Big[& Y_{i,j,n} \log \sigma(\hat{Y}_{i,j,n}) \\ & + (1 - Y_{i,j,n}) \log\big(1 - \sigma(\hat{Y}_{i,j,n})\big)\Big], \end{aligned} \tag{13}$$

where $\hat{Y}_{i,j,n}$ denotes the predicted logits, $Y_{i,j,n} \in \{0,1\}$ represent the ground-truth labels and mask $M_{i,j} \in \{0,1\}$, $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function.

3 Experiments

3.1 Datasets and Evaluation Metric

Our experiments adhere to established OVSS protocols (Ghiasi et al. 2022; Liang et al. 2023; Cho et al. 2024; Xie et al. 2024). The model is trained on COCO-Stuff (Caesar,

Method	VLM	Feature backbone	Training Dataset	Additional Dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
GroupViT (Xu et al. 2022a)	ViT-S/16	-	GCC+YFCC	✓	4.3	4.9	10.6	25.9	50.7	-
ZegFormer (Ding et al. 2022)	ViT-B/16	ResNet-101	COCO-Stuff-156	✗	4.9	9.1	16.9	42.8	86.2	62.7
ZSseg (Xu et al. 2022b)	ViT-B/16	ResNet-101	COCO-Stuff	✗	7.0	-	20.5	47.7	88.4	-
OpenSeg (Ghiasi et al. 2022)	ALIGN	ResNet-101	COCO Panoptic	✓	4.4	7.9	17.5	40.1	-	63.8
DeOP (Han et al. 2023)	ViT-B/16	ResNet-101c	COCO-Stuff-156	✗	7.1	9.4	22.9	48.8	91.7	-
PACL (Mukhoti et al. 2023)	ViT-B/16	-	GCC+YFCC	✓	-	-	31.4	50.1	72.3	-
OVSeg (Liang et al. 2023)	ViT-B/16	ResNet-101c	COCO-Stuff+COCO Caption	✓	7.1	11.0	24.8	53.3	92.6	-
SAN (Xu et al. 2023b)	ViT-B/14	-	COCO-Stuff	✗	10.1	12.6	27.5	53.8	94.0	-
CAT-Seg (Cho et al. 2024)	ViT-B/16	ResNet-101	COCO-Stuff	✗	8.4	16.6	27.2	57.5	93.7	<u>78.3</u>
EBSeg (Shan et al. 2024)	ViT-B/16	-	COCO-Stuff	✗	11.1	17.3	30.0	56.7	94.6	-
SED (Xie et al. 2024)	ConvNeXt-B	-	COCO-Stuff	✗	11.4	18.6	31.6	57.3	94.4	-
LoGoSeg (ours)	ConvNeXt-B	-	COCO-Stuff	✗	<u>12.2</u>	<u>19.3</u>	<u>32.0</u>	<u>57.8</u>	<u>95.0</u>	77.2
	ViT-B/16	-	COCO-Stuff	✗	**12.6**	**19.5**	**32.6**	**58.2**	**95.4**	**78.6**
LSeg (Li et al. 2022)	ViT-B/32	ViT-L/16	PASCAL VOC-15	✗	-	-	-	-	52.3	-
OpenSeg (Ghiasi et al. 2022)	ALIGN	Eff-B7	COCO Panoptic	✓	8.1	11.5	26.4	44.8	-	70.2
OVSeg (Liang et al. 2023)	ViT-L/14	Swin-B	COCO-Stuff+COCO Caption	✓	9.0	12.4	29.6	55.7	94.5	-
ODISE (Xu et al. 2023a)	ViT-L/14	-	COCO Panoptic	✓	11.1	14.5	29.9	57.3	-	-
HIPIE (Wang et al. 2023)	BERT-B	ViT-H	COCO Panoptic	✓	-	-	29.0	59.3	-	-
SAN (Xu et al. 2023b)	ViT-L/14	-	COCO-Stuff	✗	13.7	17.1	33.3	60.2	95.5	-
CAT-Seg (Cho et al. 2024)	ViT-L/14	Swin-B	COCO-Stuff	✗	10.8	20.4	31.5	62.0	96.6	81.8
FC-CLIP (Yu et al. 2023)	ConvNeXt-L	-	COCO Panoptic	✓	14.8	18.2	34.1	58.4	95.4	-
EBSeg (Shan et al. 2024)	ViT-L/14	-	COCO-Stuff	✗	13.7	21.0	32.8	60.2	96.4	-
SED (Xie et al. 2024)	ConvNeXt-L	-	COCO-Stuff	✗	13.9	22.6	35.2	60.6	96.1	-
MAFT+ (Jiao et al. 2024)	ConvNeXt-L	-	COCO-Stuff	✗	15.1	21.6	36.1	59.4	96.5	-
LoGoSeg (ours)	ConvNeXt-L	-	COCO-Stuff	✗	<u>15.8</u>	<u>23.6</u>	<u>37.8</u>	<u>63.2</u>	<u>97.0</u>	<u>82.1</u>
	ViT-L/14	-	COCO-Stuff	✗	**16.4**	**24.5**	**38.2**	**63.6**	**97.4**	**83.3**

Table 1: **Comparison with state-of-the-art methods.** OVSS performance on six benchmarks, reported in mIoU (%). Bold entries denote the best results, and underlined entries indicate the second best.

Uijlings, and Ferrari 2018), which comprises 118K images annotated with 171 semantic categories. For comprehensive evaluation, we test on three widely-used benchmarks: ADE20K (Zhou et al. 2019), PASCAL VOC (Everingham et al. 2010), and PASCAL-Context (Mottaghi et al. 2014).

ADE20K. The dataset provides 20 K training and 2 K validation images. We evaluate two splits: A-150, comprising the 150 most frequent categories, and A-847, which includes 847 categories following the split of (Ding et al. 2022).

PASCAL VOC. The dataset contains 1.5 K training and 1.5 K validation images annotated with 20 foreground object classes and a background label. We report results for the PAS-20 setting (foreground classes only) and for PAS-20^b, where the background is redefined to include categories from PC-59 that are not in PAS-20, as in (Ghiasi et al. 2022).

PASCAL-Context. This extension of PASCAL VOC offers dense pixel-level annotations. We follow the standard PC-59 and PC-459 splits, covering 59 and 459 semantic categories, respectively, to evaluate both common and fine-grained concepts.

Results are reported as the mean Intersection-over-Union (mIoU), a metric that averages IoU over all categories and is standard in segmentation benchmarks. Higher mIoU indicates better performance.

3.2 Implementation Details

Our framework leverages CLIP's Vision Transformer-based image encoder while keeping the text encoder frozen during training. All experiments were conducted on a $4\times$ NVIDIA RTX 4090 GPU setup with a total batch size of 4 across roughly 80K training iterations.

We employ the AdamW (Loshchilov and Hutter 2017) optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. To stabilize training, the learning rate of the image encoder is scaled by a factor of 0.01. A warm-up cosine learning rate schedule is used to gradually adjust the learning rate over time.

Training images are first cropped and resized during preprocessing to match the input resolution requirements of the CLIP backbone. To improve model robustness, we apply standard data augmentation techniques, including random horizontal flipping during training. During inference, images are uniformly resized and segmented using a sliding-window strategy to ensure full spatial coverage. For text guidance, we construct category prompts using the template *"A photo of a {category} in the scene"*. All experiments are implemented in PyTorch (Paszke et al. 2019) and built upon the Detectron2 framework (Wu et al. 2019).

3.3 Comparison With State-of-the-art Methods

We extensively evaluate our proposed LoGoSeg against leading OVSS methods on six standard benchmarks—A-847, PC-459, A-150, PC-59, PAS-20, and PAS-20^b (see Table 1). Our evaluation groups methods by type and scale of their vision-language models and provides detailed annotations on training data, external datasets, and supplementary feature backbones. The best and second-best performances

are highlighted in bold and underlined.

Most approaches rely on ViT backbones and often use extra data (e.g., COCO Captions (Chen et al. 2015), YFCC (Thomee et al. 2016)) or additional feature extractors (e.g., ResNet-101 (He et al. 2016), Swin-B (Liu et al. 2021)). In contrast, LoGoSeg employs a unified VLM backbone without external datasets or auxiliary modules. Moreover, our updated design supports diverse backbone configurations beyond the original ViT-B/16 and ViT-L/14 settings.

Backbone Scaling. Table 1 also reveals the scalability of different methods when transitioning from medium-size (e.g., ViT-B/16, ConvNeXt-B) to large backbones (ViT-L/14, ConvNeXt-L). Most two-stage frameworks show marginal gains, whereas LoGoSeg achieves near-linear improvements, consistently ranking first or second on all benchmarks (after scaling). This confirms the robustness of our object prior, region alignment, and dual-stream fusion across model capacities.

Backbone Configurations and Scaling. Using a medium backbone (ViT-B/16), LoGoSeg reaches 12.6 mIoU on A-847 and 19.5 on PC-459. When switching to the larger ViT-L/14 backbone, performance consistently improves, achieving 16.4 on A-847, 24.5 on PC-459, 38.2 on A-150, 63.6 on PC-59, 97.4 on PAS-20, and 83.3 on PAS-20^b. These results indicate a clear and consistent scaling trend. ConvNeXt-B and ConvNeXt-L variants (Liu et al. 2022) exhibit similar behavior and match or surpass their ViT counterparts. All configurations operate without external data or auxiliary modules, confirming that increasing backbone capacity translates almost linearly into accuracy gains for LoGoSeg.

One-stage vs. Two-stage Frameworks. Two-stage models like ZegFormer, OVSeg, FC-CLIP, and MAFT+ separate mask generation and classification, whereas one-stage models (SAN, CAT-Seg, SED, EBSeg) predict masks directly but emphasise either local or global cues. LoGoSeg unifies both cues in a single stage, achieving the best or second-best mIoU on all six benchmarks, as shown in Table 1. It outperforms the strongest two-stage competitor (MAFT+) by 4.2 mIoU on PC-59 and the best one-stage baseline (SED) by 1.9 mIoU on PC-459. LoGoSeg is the only one-stage method that leading on both ADE20K splits (A-150 and A-847), highlighting its balanced capability for common and long-tail categories. These results confirm that our object prior, region-aware alignment, and dual-stream fusion remain effective across backbone scales and framework styles.

Fig. 3 shows qualitative comparisons. The first row presents ground truth annotations, the second row displays SED predictions, and the third row shows LoGoSeg results. Thanks to region-aware alignment and dual-branch fusion, our method yields more stable segmentation with fewer hallucinations, correctly identifying objects like *pizza* (col. 2) and *cat* (col. 3), while avoiding misclassifications by SED. Failure cases reveal that LoGoSeg may under-segment small or heavily occluded objects (e.g., the wine glass in col. 2 and the potted plant in col. 5) and confuse visually similar classes under low-contrast conditions. See the Appendix for additional qualitative results.

OP	CCF	RA	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
			10.8	17.6	28.6	53.9	94.3	73.1
✓			11.5	18.1	29.4	56.4	94.6	75.8
✓	✓		12.3	19.0	32.0	57.8	94.8	76.9
✓	✓	✓	**12.6**	**19.5**	**32.6**	**58.2**	**95.4**	**78.6**

Table 2: **Ablation study of core modules.** Object Prior (OP), Contextual Crossmodal Fusion (CCF), and Region Alignment (RA) are evaluated using CLIP (ViT-B/16). Results are in mIoU (%).

Method	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
w/ LQ only	10.9	17.8	29.4	55.5	93.2	75.3
w/ LQ + GC	11.8	18.4	31.2	56.7	94.6	76.8
w/ LQ + GC + DA	**12.6**	**19.5**	**32.6**	**58.2**	**95.4**	**78.6**
w/ fusion depth = 1	11.5	18.5	31.6	56.9	94.4	76.6
w/ fusion depth = 2	**12.6**	**19.5**	**32.6**	**58.2**	**95.4**	**78.6**
w/ fusion depth = 3	11.9	18.8	31.5	57.4	94.8	77.2

Table 3: **Ablation of Contextual Crossmodal Fusion.** We analyze the contribution of LangQuery (LQ), GlobalContext (GC), and Dir-Attn (DA), together with different fusion depths. Results are shown in mIoU (%).

Overall, LoGoSeg demonstrates robust performance across diverse segmentation benchmarks without relying on additional data, handcrafted prompts, or auxiliary modules.

3.4 Ablation Study

We conduct ablation studies to evaluate the effectiveness of each module in LoGoSeg. All experiments are conducted under ViT-B/16 unless stated otherwise.

Impact of Core Modules. Table 2 shows the impact of integrating Object Prior (OP), Contextual Cross-modal Fusion (CCF), and Region Alignment (RA). The baseline uses global features for pixel-level segmentation without explicit local or regional alignment, yielding relatively lower scores. Incorporating OP notably reduces classification hallucinations, improving performance on all benchmarks. Adding CCF further enhances local-global feature integration, resulting in consistent performance gains. Integrating RA provides the most substantial improvement, demonstrating its effectiveness in enhancing pixel-text alignment.

Contextual Cross-modal Fusion Analysis. Table 3 evaluates the contributions of individual components in the fusion module. LangQuery alone already improves performance, indicating the effectiveness of basic text-guided attention. Incorporating GlobalContext further enhances accuracy across all benchmarks, highlighting the role of long-range semantic modeling. Directional attention brings additional gains, particularly on fine-grained datasets, due to improved sensitivity to spatial structure. Analysis of fusion depth shows that a two-layer design achieves a favorable balance between performance and complexity. Overall, these results underscore the importance of jointly modeling spatial and semantic information.

Prior-Guided Alignment Hyperparameters. We examine the effects of two hyperparameters in Table 4: the seman-

Figure 3: **Qualitative comparison.** Rows show ground truth, SED predictions, and LoGoSeg results. Region-aware alignment and dual-branch fusion yield more stable, less hallucinatory segmentations. LoGoSeg correctly identifies *pizza* and *cat* missing from the ground truth and avoids SED's mislabels (e.g., vegetables as *banana*, paintings as *TV*).

λ	$r \times r$	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
0	4×4	11.6	18.8	31.6	56.9	94.5	76.4
0	6×6	11.8	18.5	31.8	57.5	94.8	76.9
0	8×8	11.5	18.5	31.4	56.6	94.9	75.8
0.5	6×6	12.0	18.9	31.9	57.5	95.2	77.4
0.5	8×8	11.4	18.3	31.2	57.3	94.8	77.0
Adapt.	6×6	12.3	18.9	32.3	57.6	95.0	76.8
Full	6×6	**12.6**	**19.5**	**32.6**	**58.2**	**95.4**	**78.6**

Table 4: **Ablation of Prior-Guided Alignment Hyperparameters.** Varying λ and region size $r \times r$, we report mIoU (%). The *Full* setting ($\lambda = 1$, 6×6 grid) performs best. "Adapt." denotes a learnable parameter.

tic prior weight (λ) and region size ($r \times r$ grid). Results show that region alignment without object priors ($\lambda = 0$) achieves limited performance, while prior guidance yields significant improvements (optimal at $\lambda = 1$). For region granularity, a 6×6 partitioning provides a balance between localized detail preservation and computational efficiency, outperforming both larger and smaller grids.

Training Dataset Generalization. Table 5 evaluates LoGoSeg under three regimes: full COCO-Stuff, PC-59 (59 categories), and A-150 (150 categories). With PC-59 supervision, LoGoSeg lifts A-847 mIoU from 3.8 to 10.1 (+6.3) and PC-459 from 8.2 to 17.3 (+9.1) over ZegFormer, demonstrating strong transfer. Under A-150 supervision, it attains 15.5 mIoU on A-847 and 16.7 on PC-459, surpassing CAT-Seg by 2.2 points. These results verify that region-aware alignment and dual-branch fusion maintain stable performance across varying annotation scales and datasets.

Methods	Training dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
ZegFormer	COCO-Stuff	5.6	10.4	18.0	45.5	89.5	65.5
ZSseg	COCO-Stuff	7.0	9.0	20.5	47.7	88.4	67.9
CAT-Seg	COCO-Stuff	8.4	16.6	27.2	57.5	93.7	78.3
LoGoSeg (Ours)	COCO-Stuff	**12.6**	**19.5**	**32.6**	**58.2**	**95.4**	**78.6**
ZegFormer	PC-59	3.8	8.2	13.1	48.7	86.5	66.8
ZSseg	PC-59	3.0	7.6	11.9	54.7	87.7	71.7
CAT-Seg	PC-59	5.6	12.9	23.0	62.4	87.3	79.0
LoGoSeg (Ours)	PC-59	**10.1**	**17.3**	**28.2**	63.6	**93.1**	**80.1**
ZegFormer	A-150	6.8	7.1	33.1	34.7	77.2	53.6
ZSseg	A-150	7.6	7.1	40.3	39.7	80.9	61.1
CAT-Seg	A-150	10.6	14.5	46.8	46.7	85.5	70.3
LoGoSeg (Ours)	A-150	**15.5**	**16.7**	47.6	**50.8**	**91.9**	**72.7**

Table 5: **Training Dataset Generalization.** Using CLIP (ViT-B/16), we compare mIoU (%) across training sets. LoGoSeg maintains strong generalization even with limited data. Training-set scores are shown in gray.

4 Conclusion

We present LoGoSeg, a single-stage framework for OVSS that integrates object priors, region-level alignment, and local-global feature fusion. Unlike existing approaches, our method establishes region-text correspondences while suppressing irrelevant categories and addressing prediction hallucination. The dual-stream fusion architecture captures fine-grained spatial structures and global semantic context. Without external datasets or auxiliary backbones, LoGoSeg generalizes across six benchmarks. It achieves strong performance, but its reliance on large vision-language backbones may constrain real-time deployment in resource-limited scenarios. Future work includes developing lightweight variants, incorporating advanced visual encoders, and supporting multimodal inputs beyond vision-language pairs to enhance scalability and applicability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62306072 and 62576093, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242025K30024.

References

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *ArXiv*, abs/1504.00325.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Cho, S.; Shin, H.; Hong, S.; Arnab, A.; Seo, P. H.; and Kim, S. 2024. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. arXiv:2303.11797.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11583–11592.

Ding, Z.; Wang, J.; and Tu, Z. 2022. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*.

Dong, X.; Bao, J.; Zheng, Y.; Zhang, T.; Chen, D.; Yang, H.; Zeng, M.; Zhang, W.; Yuan, L.; Chen, D.; et al. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10995–11005.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, 540–557. Springer.

Han, C.; Zhong, Y.; Li, D.; Han, K.; and Ma, L. 2023. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1086–1096.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Jiao, S.; Zhu, H.; Huang, J.; Zhao, Y.; Wei, Y.; and Shi, H. 2024. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, 399–416. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *Proceedings of the International Conference on Learning Representations*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7061–7070.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7086–7096.

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

Mukhoti, J.; Lin, T.-Y.; Poursaeed, O.; Wang, R.; Shah, A.; Torr, P. H.; and Lim, S.-N. 2023. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19413–19423.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Shan, X.; Wu, D.; Zhu, G.; Shao, Y.; Sang, N.; and Gao, C. 2024. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28412–28421.

Srivastava, S.; and Sharma, G. 2024. Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 27412–27424.

Sun, L.; Cao, J.; Xie, J.; Jiang, X.; and Pang, Y. 2025. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23199–23209.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2): 64–73.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Wang, X.; Li, S.; Kallidromitis, K.; Kato, Y.; Kozuka, K.; and Darrell, T. 2023. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36: 21429–21453.

Wang, Z.; Feng, T.; Lyu, F.; Shang, F.; Feng, W.; and Wan, L. 2025. Dual Semantic Guidance for Open Vocabulary Semantic Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20212–20222.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2. Accessed: November 8, 2025.

Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; and Akata, Z. 2019. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8256–8265.

Xie, B.; Cao, J.; Xie, J.; Khan, F. S.; and Pang, Y. 2024. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3426–3436.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022a. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18134–18144.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023a. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023b. SAN: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Xu, X.; Xiong, T.; Ding, Z.; and Tu, Z. 2023c. MasQ-CLIP for Open-Vocabulary Universal Image Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 887–898.

Yang, Y.; Deng, J.; Li, W.; and Duan, L. 2025. ResCLIP: Residual Attention for Training-free Dense Vision-language Inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29968–29978.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2023. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36: 32215–32234.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, D.; Liu, F.; and Tang, Q. 2025. CorrCLIP: Reconstructing Patch Correlations in CLIP for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 24677–24687.

Zhao, H.; Puig, X.; Zhou, B.; Fidler, S.; and Torralba, A. 2017. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2002–2010.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

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

Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11175–11185.