# Bridging Naive Semantics through Proxy Injection and Matrix Fusion for Open-Vocabulary Semantic Segmentation

Zicheng Wang[1], Xiushan Nie[1,2], Yang Ning[1,2,*], Jingyuan Fang[1], Runhu Zhao[1]
[1]School of Computer and Artificial Intelligence, Shandong Jianzhu University, Jinan, P.R. China
[2]Shandong Key Laboratory of Static Traffic Large Model
[*]Corresponding author
Emails: 2023115018@stu.sdjzu.edu.cn, niexsh@hotmail.com, ningyang20@sdjzu.edu.cn,
2022110105@stu.sdjzu.edu.cn, 2023115001@stu.sdjzu.edu.cn

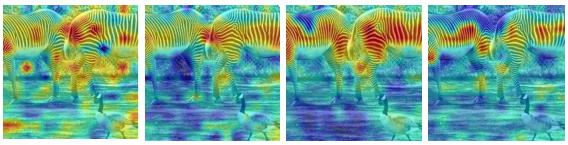
***Abstract*—Open-Vocabulary Semantic Segmentation (OVSS) aims to achieve pixel-level classification under arbitrary categories. Although recent methods leverage foundation models like SAM to enhance segmentation capabilities, weakly supervised OVSS still faces three critical challenges: (1) intermediate-layer semantic compression that loses fine-grained noun distinctions, (2) rigid feature fusion strategies that fail to adaptively balance semantic and boundary cues across diverse scenes, and (3) single-point semantic attraction where models attend only to the most salient instance when multiple objects share the same noun. To address these issues, we propose BNS-Seg, bridging naive semantics through proxy injection and matrix fusion for open-vocabulary segmentation. First, Semantic Proxy Injection (SPI) selects proxy tokens from attention key/value pairs and injects noun-level semantics into intermediate visual layers via differential attention, enhancing fine-grained semantic discrimination. Second, Doubly Stochastic Matrix Fusion (DSF) predicts sample-adaptive fusion weights under doubly stochastic constraints, enabling flexible yet stable integration of CLIP semantic features and SAM boundary features. Third, Peak-Anchored Query Diversification (PQD) identifies multiple local peaks on text-conditioned similarity maps as spatial anchors, generating diversified queries to enable effective multi-instance segmentation under the same noun. Extensive experiments on six benchmark datasets demonstrate that BNS-Seg achieves state-of-the-art performance, with improvements of up to 3.6% mIoU over existing methods.**

***Index Terms*—Open-Vocabulary; semantic segmentation; vision-language large model; cross-modal alignment; few-shot learning**

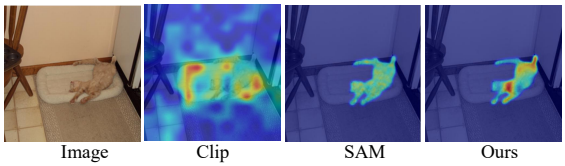
## I. INTRODUCTION

Semantic segmentation assigns a label to every pixel and supports applications such as autonomous driving, medical imaging, and remote sensing. Still, dense annotation is expensive, and fully supervised models tied to fixed label sets struggle to generalize to unseen concepts [1]. This has made text-supervised open-vocabulary semantic segmentation increasingly appealing. With vision-language models such as CLIP [2], segmentation can be linked to textual concepts rather than a closed set of predefined classes.

Early OVSS methods such as GroupViT [3] group local visual tokens and match them with text embeddings. The difficulty is that training is still driven mainly by global image-text alignment, while inference depends on much finer regional evidence. Once this gap appears, detailed parsing becomes unreliable. Later methods try to ease the mismatch from different directions. CoDe [4] introduces region-word alignment, and SynSeg [5] improves semantic discrimination through multi-class contrastive learning and feature collaboration. Another practical direction is to combine CLIP semantics with SAM priors. These advances help, but the supervision is still coarse, so weakly supervised OVSS often remains unsatisfactory at the pixel level.



(a) Semantic information decreases with increasing layer depth (from early to late stages)




(b) Complementary focuses of different models.


(c) Semantic bias toward salient instances within the same category

Fig. 1. Limitations of prior methods.

This weakness usually appears in three forms. Intermediate visual layers tend to lose subtle noun semantics, as shown in Fig. 1(a), so the model may recognize "a person riding a horse" while still confusing the rider's legs with the horse's legs. Feature fusion is also often too rigid, as illustrated in Fig. 1(b). CLIP is strong at semantics and SAM is strong at boundaries, yet a fixed fusion rule rarely fits every image. In addition, one noun may refer to several instances in the same

scene, as in Fig. 1(c). Under a single text query, the model usually focuses on the most salient object and misses the rest.

To address these issues, we propose BNS-Seg (Bridging Naive Semantics through Proxy Injection and Matrix Fusion for Open-Vocabulary Semantic Segmentation), which comprises three core modules: i) Semantic Proxy Injection (SPI): Proxy tokens are selected from attention key/value pairs, and differential attention injects noun-level semantics into intermediate layers of the visual backbone, strengthening fine-grained semantic perception. ii) Doubly Stochastic Matrix Fusion (DSF): A sample-wise doubly stochastic fusion matrix adaptively integrates CLIP semantics and SAM boundary features. The doubly stochastic constraint enables dynamic weighting while preventing any stream from being overly suppressed, and the fusion matrix itself remains interpretable. iii) Peak-Anchored Query Diversification (PQD): Multiple peaks are identified on the text-conditioned similarity map as spatial anchors, driving query diversification to enable efficient segmentation of multiple instances under the same noun and alleviating single-point semantic attraction. Our contributions are threefold:

- We identify that intermediate-layer semantic compression is a key bottleneck in weakly supervised OVSS, and address it through proxy-based differential attention injection.
- We show that fixed-weight CLIP–SAM fusion is suboptimal across diverse scenes, and propose a doubly stochastic formulation for interpretable, sample-adaptive feature integration.
- We formalize the single-point attraction problem in text-conditioned segmentation and resolve it via peak-anchored query diversification, enabling same-noun multi-instance parsing.

## II. RELATED WORK

Vision-language pretraining models such as CLIP [2] learn cross-modal representations from large-scale image-text pairs, enabling zero-shot classification and retrieval. However, their global contrastive supervision is insufficient for pixel-level dense prediction, as it lacks explicit local alignment. To address spatial modeling limitations, promptable segmentation models such as SAM [6] provide strong boundary awareness and class-agnostic mask proposals, offering complementary visual priors for open-vocabulary segmentation. Our work leverages both CLIP's semantic understanding and SAM's spatial modeling through adaptive fusion.

Existing OVSS methods can be broadly categorized into three paradigms: (1) semi-supervised approaches relying on partial pixel annotations, (2) training-free methods that directly transfer pretrained features, and (3) text-supervised segmentation using only image-caption pairs.

The text-supervised direction is further divided into global and regional alignment strategies. Global alignment methods (e.g., GroupViT [3], SimSeg [7]) match image-level features with text embeddings but suffer from train-test granularity mismatch. Regional alignment methods (e.g., TCL [8], CoDe [4]) improve fine-grained segmentation via region-text matching, yet still struggle with intermediate-layer semantic compression and single-noun multi-instance ambiguity. Our method addresses these limitations through proxy-based semantic injection and peak-anchored query diversification.

## III. THE PROPOSED METHOD

This section presents the overall pipeline of BNS-Seg, shown in Fig. 2. Given an image and its caption, we first extract nouns from the text and encode them with the CLIP text encoder. These noun embeddings guide the Semantic Proxy Injection (SPI) module, which enriches intermediate visual layers with finer semantic cues. In parallel, the image is processed by the CLIP and SAM image encoders to produce complementary semantic and boundary features. The Doubly Stochastic Matrix Fusion (DSF) module then predicts a sample-adaptive fusion matrix, normalized by Sinkhorn iterations, to combine the two streams. The fused features are matched with the noun embeddings to form text-conditioned similarity maps. Based on these maps, the Peak-Anchored Query Diversification (PQD) module selects multiple local peaks as spatial anchors and generates diversified queries, which are sent to the SAM mask decoder for final mask prediction.

### A. Semantic Proxy Injection

Under weak supervision, intermediate visual features are often too coarse. They can separate foreground from background, yet still confuse nearby nouns. For example, in an image containing "dog" and "frisbee," ambiguous pixels may still be assigned incorrectly. SPI addresses this issue by injecting noun semantics into intermediate layers. The module selects proxy tokens from attention key and value features, aligns them with noun embeddings, and writes the aligned semantics back into the visual representation.

Given a noun list $\mathcal{N} = \{n_i\}_{i=1}^{N}$, we generate prompt sentences using a predefined template set and encode them with a frozen text encoder. The text prototype for each noun is obtained by averaging over templates as $\bar{\mathbf{e}}_i = \frac{1}{M}\sum_{m=1}^{M} \mathrm{norm}(E_t(t_i^m))$, where $E_t(\cdot)$ denotes the text encoder and $t_i^m$ is the $m$-th template sentence for noun $n_i$. We stack all prototypes into a matrix $\mathbf{E} = [\bar{\mathbf{e}}_1, \ldots, \bar{\mathbf{e}}_N]^\top \in \mathbb{R}^{N\times d}$ and project it into the visual space for alignment and proxy generation.

At layer $\ell$, we compute the affinity matrix between normalized keys and text prototypes as $\mathbf{A}^\ell = \hat{\mathbf{K}}^\ell \hat{\mathbf{E}}^\top$, where $\hat{\mathbf{K}}^\ell$ and $\hat{\mathbf{E}}$ denote L2-normalized keys and text prototypes, respectively. When needed, the affinity can be modulated by a Sinkhorn transport plan to suppress overly dominant responses.

Proxy selection follows two steps. Top-$k$ retains the tokens most correlated with each noun, while Bottom-$q$ keeps low-response tokens that often lie near semantic boundaries and carry useful discriminative cues. Duplicate indices are removed to avoid redundancy.

Proxy tokens then aggregate both text and visual evidence through positive-mask pooling. We pool the positive regions

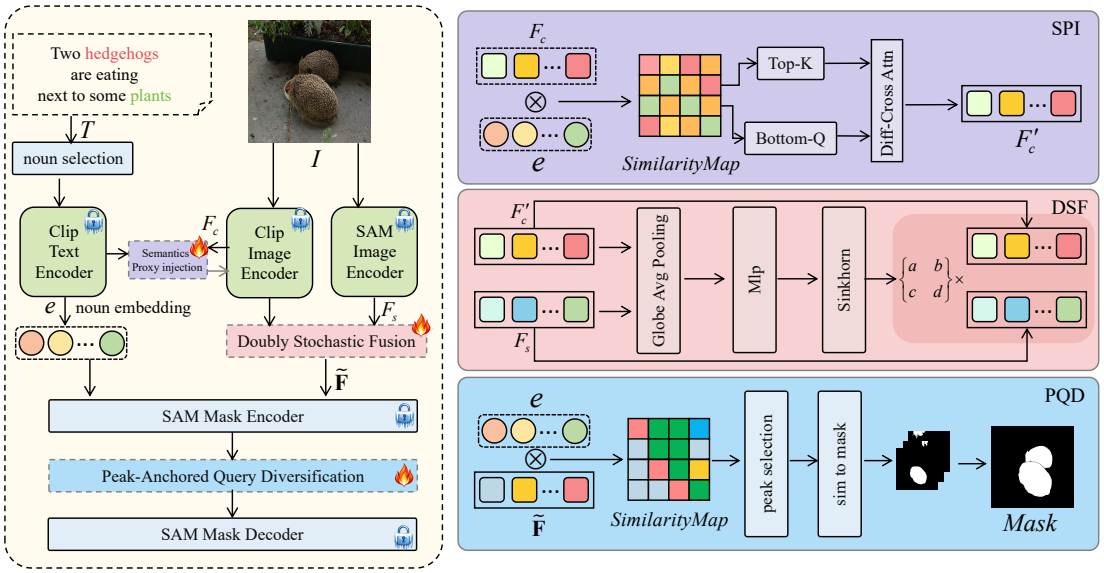



Fig. 2. Overview of the proposed BNS-Seg module.

and combine text-derived and visual-derived components with a learnable scaling factor in exponential difference form for stable training.

Standard softmax attention tends to spread weights over noisy positions. We therefore adopt differential attention:

$$\begin{aligned}\mathrm{DiffAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \big(&\mathrm{softmax}(\mathbf{Q}_1\mathbf{K}_1^\top/\sqrt{d}) \\ &- \lambda\,\mathrm{softmax}(\mathbf{Q}_2\mathbf{K}_2^\top/\sqrt{d})\big)\mathbf{V}.\end{aligned} \tag{1}$$

Proxy write-back is performed in two stages. First, proxy tokens $\mathbf{X}^\ell$ absorb text semantics by attending to text prototypes $\mathbf{E}$:

$$\mathbf{X}'^\ell = \mathbf{X}^\ell + \mathrm{DiffAttn}(\mathbf{X}^\ell, \mathbf{E}, \mathbf{E}). \tag{2}$$

The updated proxies are then written into the visual representation $\mathbf{F}_v^\ell$:

$$\tilde{\mathbf{F}}_v^\ell = \mathbf{F}_v^\ell + \alpha \cdot \mathrm{DiffAttn}(\mathbf{F}_v^\ell, \mathbf{X}^\ell, \mathbf{X}'^\ell). \tag{3}$$

Here $\alpha$ is initialized to $10^{-3}$ so that semantic injection does not disturb visual features too strongly at the beginning of training.

To align proxy tokens with noun embeddings, we use a bidirectional contrastive loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\big(\mathrm{CE}(\mathrm{softmax}(\mathbf{S}/\tau), \mathbf{I}) + \mathrm{CE}(\mathrm{softmax}(\mathbf{S}^\top/\tau), \mathbf{I})\big), \tag{4}$$

where $\mathbf{S} = \mathbf{X}^\ell\mathbf{E}^\top$ is the similarity matrix between proxy tokens and text prototypes, $\tau$ is a learnable temperature, and $\mathbf{I}$ is the identity matrix. This loss encourages one-to-one correspondence between proxy tokens and noun embeddings.

### *B. Doubly Stochastic Matrix Fusion*

CLIP provides strong semantics, whereas SAM offers cleaner boundaries. A fixed fusion rule cannot suit all images, but unconstrained sample-wise weights can easily oversuppress one branch and destabilize optimization. We therefore introduce the Doubly Stochastic Matrix Fusion (DSF) module.

For each input image, DSF predicts a $2 \times 2$ fusion matrix $\mathbf{M} \in \mathbb{R}^{2\times 2}$ satisfying the doubly stochastic constraints:

$$M_{ij} \geq 0, \quad \sum_j M_{ij} = 1, \quad \sum_i M_{ij} = 1. \tag{5}$$

These constraints keep the fusion adaptive but controlled. The weights vary with the input, information reduced in one output is preserved in the other, and neither feature stream is pushed to an extreme. In this way, DSF balances flexibility and stability without manual tuning.

Given CLIP and SAM feature maps aligned to the same channel dimension $C$, we first apply global average pooling to obtain compact representations:

$$\mathbf{z}_c = \mathrm{GAP}(\mathbf{F}_c), \quad \mathbf{z}_s = \mathrm{GAP}(\mathbf{F}_s), \quad \mathbf{z}_c, \mathbf{z}_s \in \mathbb{R}^{B\times C}. \tag{6}$$

The pooled features are concatenated and processed by a lightweight MLP to produce a raw score matrix $\mathbf{S}$.

To obtain a doubly stochastic matrix, we first enforce nonnegativity through exponentiation:

$$\mathbf{A}^{(0)} = \exp\big(\mathbf{S} - \max(\mathbf{S})\big). \tag{7}$$

We then apply the Sinkhorn-Knopp algorithm [9], which alternates between row and column normalization:

$$\mathbf{A} \leftarrow \mathbf{A} \oslash (\mathbf{A}\mathbf{1} + \varepsilon), \quad \mathbf{A} \leftarrow \mathbf{A} \oslash (\mathbf{1}^\top\mathbf{A} + \varepsilon), \tag{8}$$

where $\mathbf{1}$ denotes an all-ones vector, $\oslash$ represents elementwise division with broadcasting, and $\varepsilon$ is a small constant for numerical stability. After $T$ iterations, we obtain the doubly stochastic matrix $\mathbf{M} = \mathbf{A}^{(T)}$.

The fusion matrix produces two complementary fused representations:

$$[\tilde{\mathbf{F}}_1, \tilde{\mathbf{F}}_2]^\top = \mathbf{M}[\mathbf{F}_c, \mathbf{F}_s]^\top. \tag{9}$$

The two outputs emphasize semantic and boundary cues in a complementary way. We then concatenate them and apply a $1 \times 1$ convolution to obtain the final fused features:

$$\tilde{\mathbf{F}} = \mathrm{Conv}_{1\times1}\big([\tilde{\mathbf{F}}_1; \tilde{\mathbf{F}}_2]\big) \in \mathbb{R}^{B\times C\times H\times W}. \tag{10}$$

This yields an adaptive combination while preserving the stability introduced by the doubly stochastic constraint.

### C. Peak-Anchored Query Diversification

In many images, one noun corresponds to several objects. Under a single-query setting, the response usually concentrates on the most salient instance and misses the rest. PQD addresses this issue by treating multiple local peaks in the similarity map as anchors for different instances and generating one shifted query for each anchor.

Given fused visual feature map $\tilde{\mathbf{F}} \in \mathbb{R}^{C\times H\times W}$ and noun embedding $\mathbf{e} \in \mathbb{R}^C$, we first construct a text-conditioned similarity map to capture the spatial distribution of the noun. Specifically, we apply $L_2$ normalization to visual and text features and compute cosine similarity:

$$s_{hw} = \langle \bar{\mathbf{f}}_{hw}, \bar{\mathbf{e}} \rangle, \quad \bar{\mathbf{f}}_{hw} = \frac{\tilde{\mathbf{F}}_{hw}}{\|\tilde{\mathbf{F}}_{hw}\|_2}, \quad \bar{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \tag{11}$$

where $\tilde{\mathbf{F}}_{hw} \in \mathbb{R}^C$ denotes the feature vector at spatial location $(h, w)$.

Selecting only the global maximum is often insufficient, because several strong responses may cluster in one local region. We therefore keep local maxima with a max-pooling-based strategy:

$$\tilde{s}_{hw} = \begin{cases} s_{hw}, & \text{if } s_{hw} = \mathrm{MaxPool}_r(\mathbf{S})_{hw}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{12}$$

where $\mathbf{S} \in \mathbb{R}^{h\times w}$ is the similarity map and $r$ is the local window size. We then take the top-$K$ peaks from $\tilde{\mathbf{S}}$ to form the anchor set $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^K$. When ambiguity is weak, this naturally reduces to single-query localization.

For each anchor $\mathbf{p}_i$, we sample a local semantic vector and use it to shift the noun query:

$$\mathbf{f}_i = \mathrm{GridSample}(\tilde{\mathbf{F}}, \mathbf{p}_i), \quad \boldsymbol{\Delta}_i = \phi(\alpha \mathbf{f}_i), \tag{13}$$

where $\phi(\cdot)$ is a projection function (identity or a lightweight MLP) and $\alpha$ is a scaling factor.

The shifted query is then defined as $\mathbf{q}_i = \mathrm{LN}(\mathbf{e} + \boldsymbol{\Delta}_i)$. This keeps the core noun semantics while nudging the query toward a specific local region.

Each shifted query is passed through the SAM mask generation module to produce a soft mask $\mathbf{m}_i$. We then take the element-wise maximum over the $K$ candidate masks to obtain the positive mask:

$$\mathbf{m}^{\mathrm{pos}} = \max_{i\in\{1,\dots,K\}} \mathbf{m}_i. \tag{14}$$

The negative mask $\mathbf{m}^{\mathrm{neg}}$ is generated in the same way using the corresponding non-matching noun query.

To encourage reasonable foreground coverage and suppress background responses, we introduce an area-prior loss:

$$\mathcal{L}_{\mathrm{area}} = |\mathrm{mean}(\mathbf{m}^{\mathrm{pos}}) - p| + \mathrm{mean}(\mathbf{m}^{\mathrm{neg}}) \tag{15}$$

where the first term encourages positive masks to cover a target area ratio $p$, and the second term suppresses spurious activations on negative masks.

## IV. EXPERIMENTS AND DISCUSSION

### A. Datasets and Evaluation Settings

We train the proposed model on image-text pairs and evaluate it on six standard semantic segmentation benchmarks.

*1) Training Datasets:* We use Conceptual Captions 3M (CC3M) [15] and Conceptual 12M (CC12M) [16], which contain 3 million and 12 million image-text pairs and are widely used for text-supervised segmentation.

*2) Evaluation Datasets:* We evaluate on PASCAL VOC 2012 [17], Pascal Context-59 [18], COCO-Stuff and COCO-Object [19], [20], Cityscapes [21], and ADE20K [22]. Official validation splits are used throughout. VOC and COCO-Object include a background class, while the remaining datasets are evaluated on labeled categories only. Following CoDe [4], we adopt the category names provided by MMSegmentation and use its post-processing pipeline.

*3) Training Configuration:* We use CLIP ViT-B/16 as the semantic backbone and SAM as the auxiliary branch. Images are resized to $224 \times 224$ with random cropping and horizontal flipping, then normalized. For each image-text pair, two nouns are randomly selected as supervision words. We optimize the model with AdamW using a base learning rate of 3.2e-4 and a minimum of 4e-5, and train on four NVIDIA 2080 Ti GPUs following the CoDe [4] setting.

### B. Comparison with Existing Methods

We compare our method with representative text-supervised segmentation approaches on six datasets, as reported in Table I. The mIoU values are taken from the original papers; GroupViT [3], CoCu [11], and TCL [8] are evaluated only on datasets without background classes, and the training datasets used by different methods are listed in the table.

BNS-Seg achieves the best mIoU on all six datasets and the highest average score of 37.0. Compared with the strongest baselines, CoDe [4] and SynSeg [5], it delivers more consistent gains across benchmarks, suggesting better handling of semantic compression and same-noun semantic neglect.

*1) Qualitative Comparison with Prior Methods:* Fig. 3 compares our method with SimSeg [7], TCL [8], and CoDe [4]. For fairness, we only include models with released weights.

SimSeg introduces more noise in low contrast or cluttered scenes such as boats, while TCL and CoDe often break thin structures and struggle when nearby instances share the same noun, as in bird wings or adjacent sheep. In contrast, our method produces cleaner regions, sharper boundaries, and more complete masks in these challenging cases. These qualitative observations are consistent with the quantitative gains in Table I.

TABLE I
PERFORMANCE COMPARISON WITH OTHER METHODS.

| Method | Publication | Training Datasets | VOC | Context | Object | Stuff | City | ADE | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| GroupViT [3] | CVPR 2022 | CC3M+CC12M+RedCaps12M | 50.4 | 23.4 | 27.5 | 15.3 | 11.1 | 9.2 | 22.8 |
| ViewCo [10] | ICLR 2023 | CC12M+YFCC14M | 52.4 | 23.0 | 23.5 | – | – | – | – |
| CoCu [11] | NeurIPS 2023 | CC3M+CC12M+YFCC14M | 51.4 | – | 22.7 | 15.2 | 22.1 | – | – |
| OVSegmentor [12] | CVPR 2023 | CC12M | 53.8 | 20.4 | 25.1 | – | – | – | – |
| TCL [8] | CVPR 2023 | CC3M+CC12M | 55.0 | 33.9 | 31.6 | 22.4 | 24.0 | 14.9 | 30.3 |
| CoDe [4] | CVPR 2024 | CC3M+CC12M | 57.7 | 30.5 | 32.3 | 23.9 | 28.9 | 16.2 | 31.6 |
| S-Seg [13] | CVPR 2025 | CC3M+CC12M | 53.2 | 27.2 | 30.3 | – | – | – | – |
| MGCA [14] | TOMM 2025 | CC3M | 53.1 | 33.7 | 31.9 | 22.0 | 24.0 | 14.5 | 29.9 |
| SynSeg [5] | AAAI 2025 | CC12M | 62.2 | 41.8 | 34.9 | 23.6 | 30.9 | – | – |
| Ours | – | CC3M+CC12M | **65.8** | **43.7** | **36.7** | **24.7** | **32.3** | **19.1** | **37.0** |

Note: "–" denotes results not reported in the original paper; averages are computed excluding missing entries.

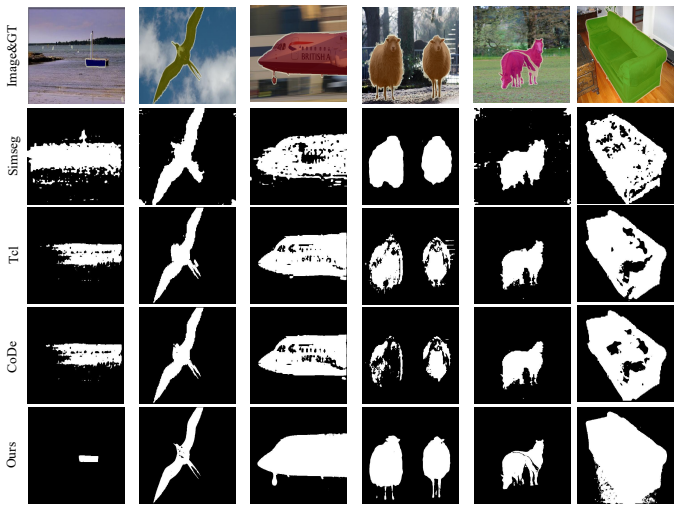



Fig. 3. Visual comparison of segmentation results.

TABLE II
ABLATION STUDY ON INDIVIDUAL COMPONENTS.

| C | W | R | VOC | Context | Object | Stuff | City | ADE |
|---|---|---|---|---|---|---|---|---|
| | | | 58.3 | 35.2 | 33.1 | 22.4 | 28.6 | 17.2 |
| ✓ | | | 62.1 | 39.5 | 34.8 | 23.5 | 30.4 | 18.1 |
| ✓ | ✓ | | 64.6 | 42.1 | 35.9 | 24.2 | 31.5 | 18.7 |
| ✓ | | ✓ | 63.4 | 40.8 | 35.3 | 23.9 | 30.9 | 18.4 |
| | ✓ | ✓ | 61.5 | 39.1 | 33.9 | 23.2 | 29.3 | 17.9 |
| ✓ | ✓ | ✓ | 65.8 | 43.7 | 36.7 | 24.7 | 32.3 | 19.1 |

*C: Semantic Proxy Injection; W: Doubly Stochastic Matrix Fusion; R: Peak-Anchored Query Diversification.*

### C. Ablation Studies

Unless otherwise stated, all ablation experiments are conducted on PASCAL VOC [17], and mIoU denotes the result on this dataset.

*1) Contribution of Each Module:* Table II reports the effect of progressively adding each component to the baseline, which consists of CLIP and SAM. Semantic Proxy Injection (C) alone improves the baseline by 3.8 mIoU on VOC and 4.3 mIoU on Context. Adding Doubly Stochastic Matrix Fusion (W) or Peak-Anchored Query Diversification (R) brings further gains, while the weaker W+R setting shows that SPI provides an essential semantic foundation. The full model achieves the best overall results, improving VOC and Context by 7.5 and 8.5 mIoU and bringing consistent gains from 1.9 to 3.7 mIoU on the remaining datasets.

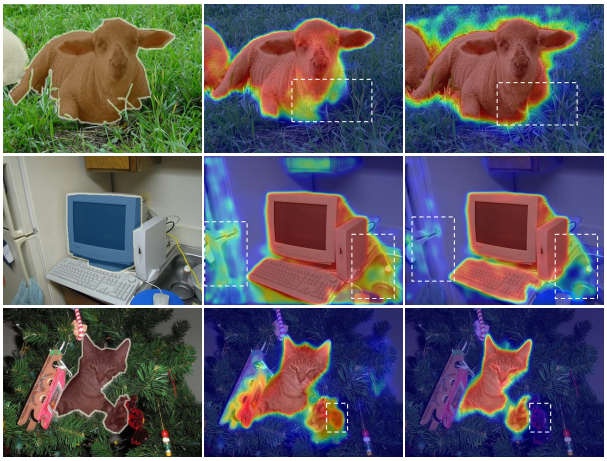



Fig. 4. Qualitative Ablation of Semantic Proxy Injection

*2) Qualitative Ablation of SPI: Reducing Spurious Activations and Enhancing Boundary Clarity:* Fig. 4 compares the settings without SPI and with SPI. Without SPI, intermediate features produce clear spurious activations in background or texture regions, such as the grass near the sheep and the clutter around the computer. With SPI, the responses become more object focused and the boundaries cleaner, especially around fine structures such as sheep legs and computer edges.

*3) Ablation results for fusion weight design choices:* Table III shows that moving from uniform fusion to a learnable scalar and then to DSF yields steady improvements, with DSF reaching the best result of 62.4 mIoU. This confirms

TABLE III
ABLATION RESULTS FOR FUSION WEIGHT DESIGN CHOICES.

| ID | Method | Description | mIoU |
|---|---|---|---|
| A | Uniform | $\tilde{\mathbf{F}} = 0.5\mathbf{F}_c + 0.5\mathbf{F}_s$ | 60.6 |
| B | Learnable Scalar | $\tilde{\mathbf{F}} = \sigma(\alpha)\mathbf{F}_c + (1-\sigma(\alpha))\mathbf{F}_s$ | 61.0 |
| C | **DSF (Ours)** | $\tilde{\mathbf{F}} = \begin{bmatrix}\alpha & \beta \\ \gamma & \delta\end{bmatrix}\begin{bmatrix}\mathbf{F}_c \\ \mathbf{F}_s\end{bmatrix}$ | **62.4** |

that structured fusion constraints improve adaptivity without sacrificing stability.

TABLE IV
ABLATION ON LAYER INJECTION DEPTH.

| Layers | [7] | **[3,7]** | [3,5,7] | [2,4,6,8] | [1,3,5,7,9] |
|---|---|---|---|---|---|
| mIoU | 64.8 | **65.8** | 65.5 | 65.1 | 64.8 |

*4) Ablation on Injection Layer Depth:* Table IV shows that injecting semantics into two intermediate layers gives the best performance. A single layer provides insufficient propagation, whereas using more layers introduces redundancy, so we adopt two layers as the default setting.

## V. CONCLUSION

We presented BNS-Seg, which addresses three fundamental limitations of prior methods. First, to mitigate intermediate-layer semantic compression, we introduced *Semantic Proxy Injection* (SPI), which selects proxy tokens from attention key/value pairs and injects noun-level semantics into the visual backbone via differential attention, enhancing fine-grained semantic discrimination. Second, to overcome rigid cross-modal fusion, we proposed *Doubly Stochastic Matrix Fusion* (DSF), which predicts a sample-adaptive fusion matrix under doubly stochastic constraints, enabling flexible yet stable integration of CLIP semantics and SAM boundary features. Third, to resolve multi-region semantic ambiguity, we designed *Peak-Anchored Query Diversification* (PQD), which identifies multiple local peaks in text-conditioned similarity maps as spatial anchors and generates diversified queries for same-noun multi-instance segmentation. This work opens new avenues for vision-language model research and their broader applications in dense prediction tasks within computer vision.

## REFERENCES


[1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *NeurIPS*, 2021.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[3] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 134–18 144.

[4] J.-J. Wu, A. C.-H. Chang, C.-Y. Chuang, C.-P. Chen, Y.-L. Liu, M.-H. Chen, H.-N. Hu, Y.-Y. Chuang, and Y.-Y. Lin, "Image-text co-decomposition for text-supervised semantic segmentation," in *CVPR*, 2024.

[5] W. Zhang, K. Liu, F. Dang, Z. Zhu, X. Sun, and Y. Liu, "Synseg: Feature synergy for multi-category contrastive learning in end-to-end open-vocabulary semantic segmentation," *arXiv preprint arXiv:2508.06115*, 2025.

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

[7] M. Yi, Q. Cui, H. Wu, C. Yang, O. Yoshie, and H. Lu, "A simple framework for text-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7071–7080.

[8] J. Cha, J. Mun, and B. Roh, "Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 165–11 174.

[9] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967.

[10] P. Ren, C. Li, H. Xu, Y. Zhu, G. Wang, J. Liu, X. Chang, and X. Liang, "Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency," *arXiv preprint arXiv:2302.10307*, 2023.

[11] Y. Xing, J. Kang, A. Xiao, J. Nie, L. Shao, and S. Lu, "Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68 798–68 809, 2023.

[12] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie, "Learning open-vocabulary semantic segmentation models from natural language supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2935–2944.

[13] Z. Lai, "Exploring simple open-vocabulary semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 221–30 230.

[14] Y. Liu, P. Ge, G. Wang, Q. Liu, and D. Huang, "Multi-grained contrastive learning for text-supervised open-vocabulary semantic segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 3, pp. 1–21, 2025.

[15] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[16] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

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

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

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

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[22] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.