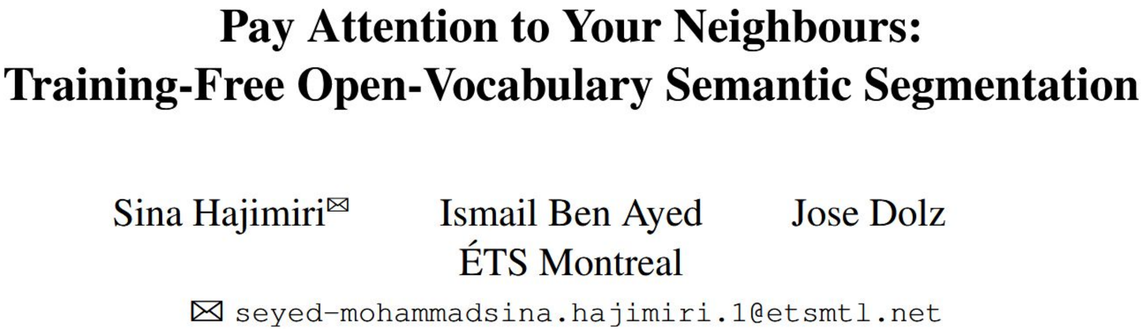
WACV 2025



Pay Attention to Your Neighbours: Training-Free Open-Vocabulary Semantic Segmentation

Sina Hajimiri✉ Ismail Ben Ayed Jose Dolz
ÉTS Montreal
✉ seyed-mohammadsina.hajimiri.1@etsmtl.net

- Problem / objective
  - Open-Vocabulary Semantic Segmentation (OVSS)
- Contribution / Key idea
  - Neighbour-Aware CLIP (NACLIP)
    - Adaptation of CLIP for OVSS
    - Training-free
    - Enforce localization of patches in the self-attention

전유진

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

WACV 2025

- **CLIP**

Vision Language model for image classification task

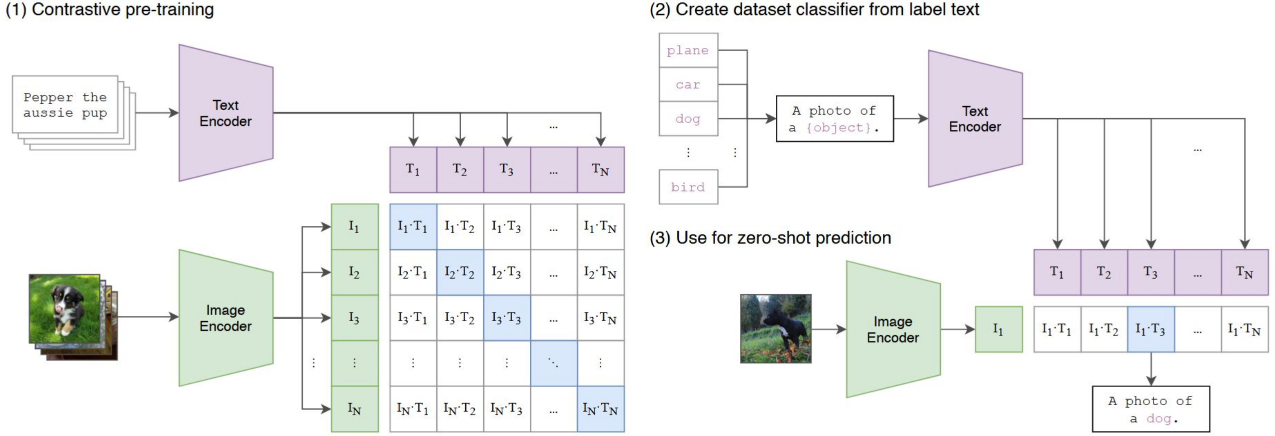

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

전유진

RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PmLR, 2021. p. 8748-8763.

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

WACV 2025

- **Limitations of adapting CLIP for Semantic Segmentation**

1. 어텐션 강도가,
해당 패치의 부근이 아닌
멀리있는 패치들에 분산됨.

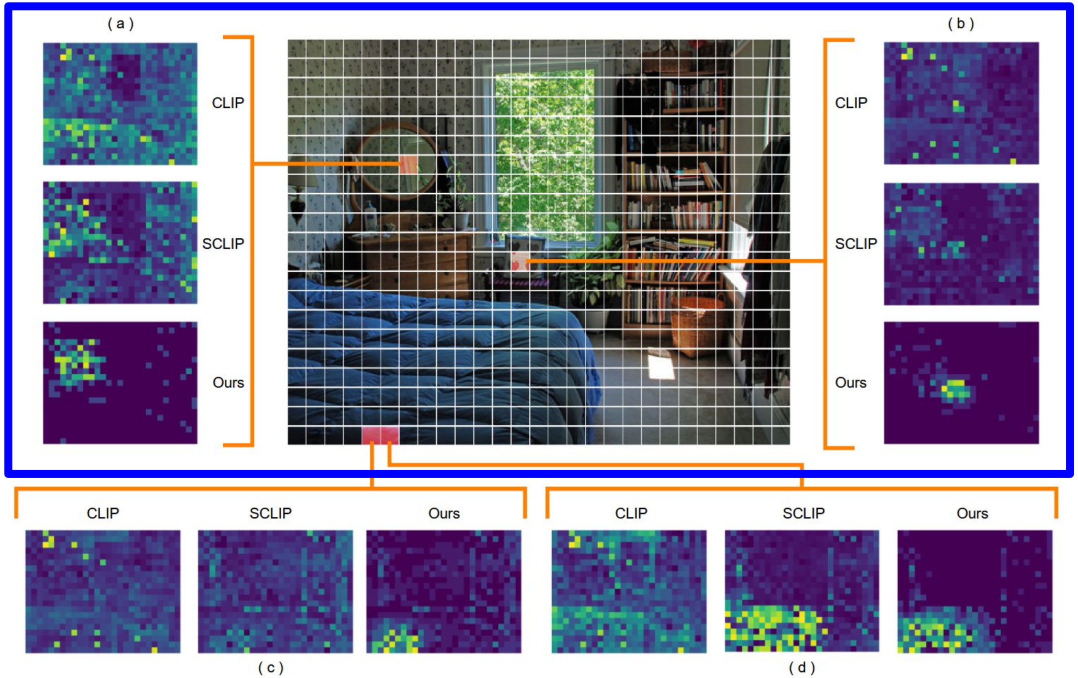


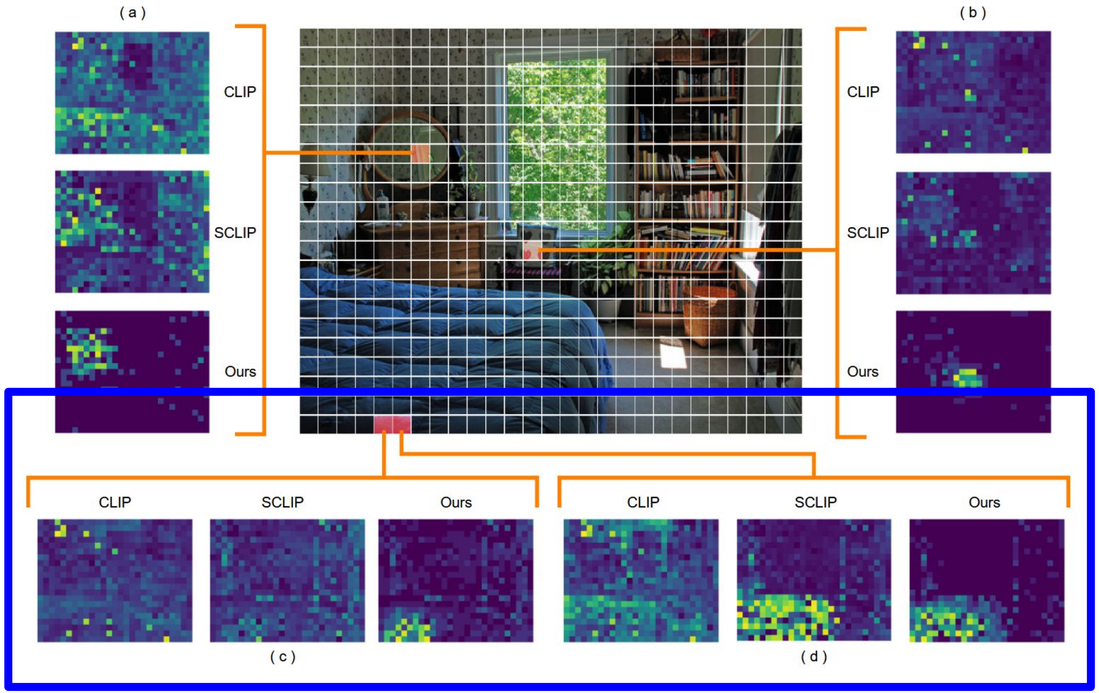
Figure 1. **Attention maps of the final visual encoder layer.** For the patches shaded in red (denoted with (a) to (d)), the final layer's attention maps are presented for CLIP [32], SCLIP [38], and our method. We have identified two problematic phenomena in the attention maps of CLIP and SCLIP, stemming from a lack of mechanisms to properly attend to patches' neighbourhoods. First, as depicted in (a) and (b), attention intensity is sometimes dispersed among distant patches, neglecting the vicinity of a patch. Additionally, adjacent or closely located patches sharing the same real-world category and even similar visual characteristics can have inconsistent attention maps. For instance, while SCLIP generates a quality attention map for patch (d), its attention map for (c) is notably different and fails to focus on the desired object. By explicitly promoting attention to neighbours, our method produces consistent attention maps across adjacent patches.

[1] WANG, Feng; MEI, Jieru; Y ... n. Cham: Springer Nature Switzerland, 2024. p. 3

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

WACV 2025

- **Limitations of adapting CLIP for Semantic Segmentation**



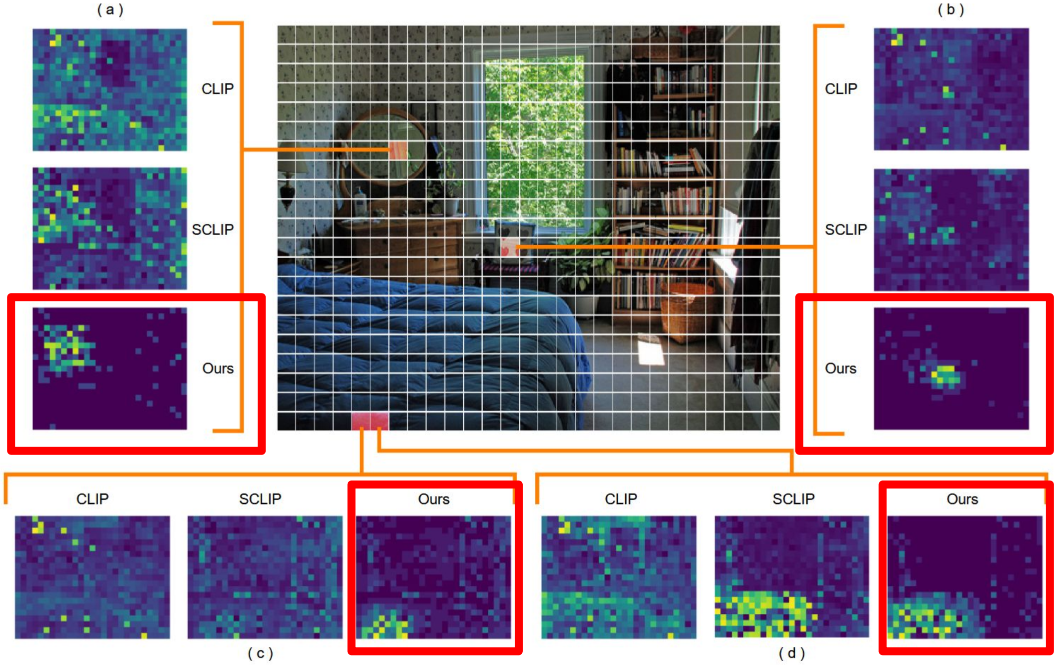
2. 동일 카테고리에 속하는 근접 패치들 간에도 attention map이 다르게 나옴

Figure 1. **Attention maps of the final visual encoder layer.** For the patches shaded in red (denoted with (a) to (d)), the final layer's attention maps are presented for CLIP [32], SCLIP [38], and our method. We have identified two problematic phenomena in the attention maps of CLIP and SCLIP, stemming from a lack of mechanisms to properly attend to patches' neighbourhoods. First, as depicted in (a) and (b), attention intensity is sometimes dispersed among distant patches, neglecting the vicinity of a patch. Additionally, adjacent or closely located patches sharing the same real-world category and even similar visual characteristics can have inconsistent attention maps. For instance, while SCLIP generates a quality attention map for patch (d), its attention map for (c) is notably different and fails to focus on the desired object. By explicitly promoting attention to neighbours, our method produces consistent attention maps across adjacent patches.

[1] WANG, Feng; MEI, Jieru; Y n. Cham: Springer Nature Switzerland, 2024. p. 3

- ## Motivation

각 패치가 공간적으로 가까운 패치들에 attend 하도록 유도하여, ***(Neighbour-Aware CLIP)***
인접한 영역에서의 attention map이 일관되도록 만듦으로서,
위치적으로 가까운 픽셀들은 동일하거나 유사한 시맨틱 클래스로 분류되게 만들겠다. ***(Local Spatial Consistency)***

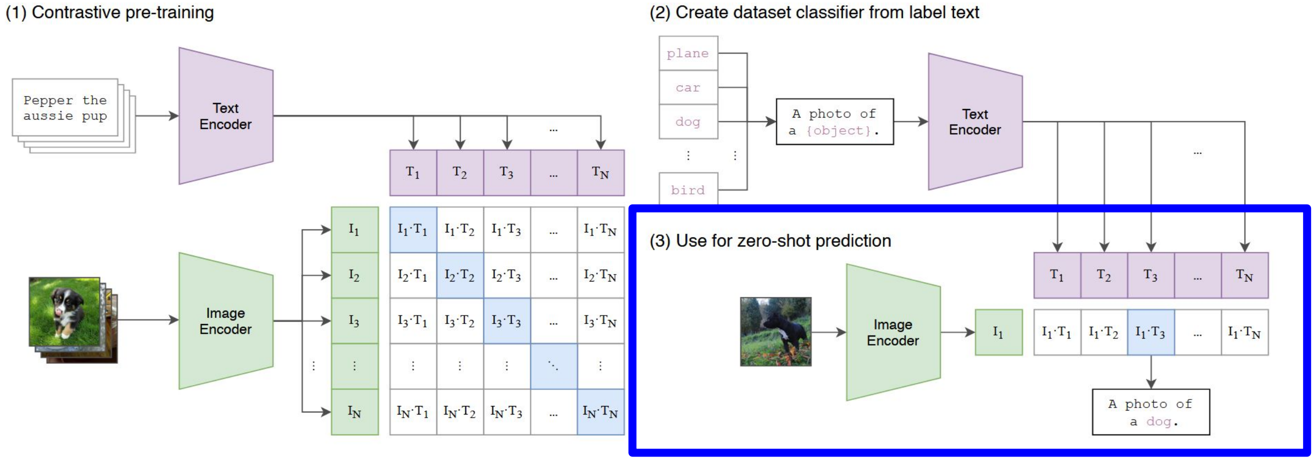


HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

WACV 2025

Preliminaries: CLIP

- **CLIP for image classification (default)**
  - compare each image representation with text representations through cosine similarity
- **CLIP for image segmentation**
  - compare each patch representation with text representations through cosine similarity



(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
A photo of a dog.

	T_1	T_2	T_3	...	T_N
I_1	$I_1{\cdot}T_1$	$I_1{\cdot}T_2$	$I_1{\cdot}T_3$	...	$I_1{\cdot}T_N$
I_2	$I_2{\cdot}T_1$	$I_2{\cdot}T_2$	$I_2{\cdot}T_3$	...	$I_2{\cdot}T_N$
I_3	$I_3{\cdot}T_1$	$I_3{\cdot}T_2$	$I_3{\cdot}T_3$	...	$I_3{\cdot}T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N{\cdot}T_1$	$I_N{\cdot}T_2$	$I_N{\cdot}T_3$	...	$I_N{\cdot}T_N$

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

전유진

RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PmLR, 2021. p. 8748-8763.

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

● Preliminaries: CLIP

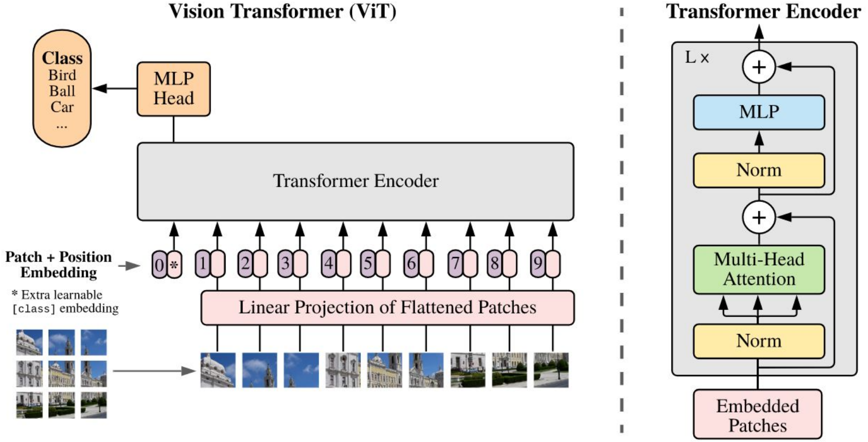


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

❑ Encoder block

$$\mathbf{Z}' = \mathrm{LN}(\mathbf{Z}^{(l-1)}), \qquad (1)$$

$$\mathbf{Z}' = \mathbf{Z}^{(l-1)} + \mathrm{SA}(\mathbf{Z}'), \qquad (2)$$

$$\mathbf{Z}^* = \mathrm{LN}(\mathbf{Z}'), \qquad (3)$$

$$\mathbf{Z}^{(l)} = \mathbf{Z}' + \mathrm{MLP}(\mathbf{Z}^*). \qquad (4)$$

❑ Self-attention module

$$[\mathbf{q},\mathbf{k},\mathbf{v}] = \mathbf{Z}\mathbf{W}^{qkv}, \qquad (5)$$

$$\mathrm{sim}_{ij} = \frac{\mathbf{kq}_{ij}}{\sqrt{d}}, \qquad (6)$$

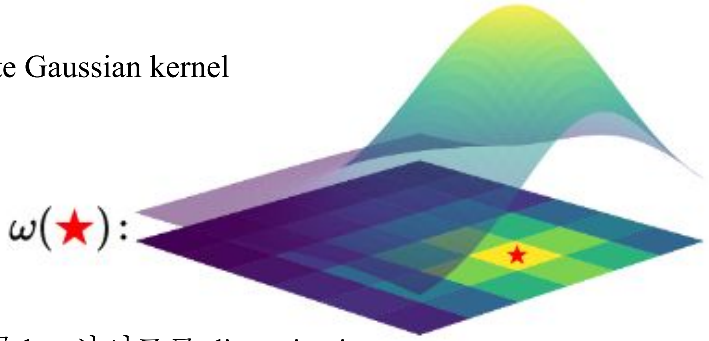
$$A_{ij} = \mathrm{softmax}\left(\mathrm{sim}_{ij}\right)\mathbf{v}, \qquad (7)$$

$$\mathrm{SA}(\mathbf{Z})_{ij} = A_{ij}\mathbf{W}^{o}. \qquad (8)$$

RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PmLR, 2021. p. 8748-8763.

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

● Neighbour-Aware CLIP: (1) Introducing spatial consistency



- ❑ Augment the attention map information with an unnormalized multivariate Gaussian kernel
- ❑ Gaussian kernel: x = μ 에서 최대이고, 멀어질수록 감소.

$$\phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \quad (9)$$

$$\phi(\mathbf{x};\boldsymbol{\mu},\sigma) = \exp\left(-\frac{||\mathbf{x}-\boldsymbol{\mu}||^2}{2\sigma^2}\right), \quad (10) \qquad \boldsymbol{\Sigma} = \sigma^2 I$$

- ❑ Discretized Gaussian kernel: (i, j) 패치를 중심으로 한 Gaussian kernel를 hw 사이즈로 discretization

$$\begin{gathered}\omega((i,j);\sigma)_{mn} = \phi((m,n);(i,j),\sigma), \\ \forall m \in \{1,2,\ldots,h\},\ \forall n \in \{1,2,\ldots,w\},\end{gathered} \quad (11)$$

- ❑ *Neighbourhood Only Attention*: attention map's logits are zero (Image-independent한데, 이것만 해도 성능 급증)

$$A_{ij} = \operatorname{softmax}\left(\omega((i,j),\sigma)\right)\mathbf{v}, \quad (12)$$

- ❑ ∴ Add the Gaussian window to the logits of the attention map for patch (i, j) (수식 7로부터 개선)

$$\widetilde{A}_{ij} = \operatorname{softmax}\left(\mathrm{sim}_{ij} + \omega((i,j);\sigma)\right)\mathbf{v}, \quad (13)$$

$$\widetilde{\mathrm{SA}}(\mathbf{Z})_{ij} = \widetilde{A}_{ij}\mathbf{W}^o. \quad (14)$$

Neighbour-Aware CLIP: (2) Measure of similarity

- ❑ *Key-Key Similarity*
  - 기존의 Query-Key similarity 대신 Key-Key similarity 사용
  - CLIP: Image classification task에 맞게 query 중심의 attention 사용
  - NACLIP:
    - i) Image segmentation task에서는 각 패치의 속성이 더 중요하므로, query의 시점이 아니라 key 간의 관계에 주목.
    - ii) 쿼리가 키에 얼마나 주목할지 → 키와 키가 얼마나 비슷한 정보를 표현하는가를 측정

$$\mathrm{sim}_{ij} = \frac{\mathbf{kq}_{ij}}{\sqrt{d}} \quad \rightarrow \quad \frac{\mathbf{kk}_{ij}}{\sqrt{d}}$$

- ❑ ∴ (수식 13으로부터 개선)

$$\widetilde{A}_{ij} = \mathrm{softmax}\left(\frac{\mathbf{kk}_{ij}}{\sqrt{d}} + \omega((i,j);\sigma)\right)\mathbf{v}. \qquad (15)$$

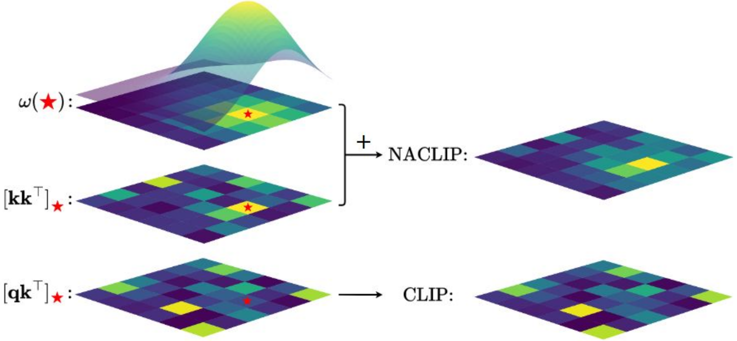


Figure 2. **Schematic figure depicting the mechanism to form attention maps.** Maps are shown for the patch located at ★, ω(★) denotes a discretized Gaussian kernel centered at ★ (see Eq. (11)), and $[\mathbf{xy}^{\top}]_{★}$ indexes $\mathbf{xy}^{\top}$ on patch ★. CLIP does not ensure high attention to the patch itself and the neighbouring patches, while NACLIP does. Scaling and softmax operations are omitted for demonstration simplicity.

- **Neighbour-Aware CLIP: (3) Eliminating image-level specialized units**

❑ Motivation

- MaskCLIP[1]: "CLIP’s vision transformer의 final encoder block이 dense prediction하기에 적합하지 않은 구조이다."
- 따라서, 우리는 CLIP이 semantic segmentation에 적합하도록, final encoder block의 구조만 변경함.

❑ Method

1. Feed-forward block 제거

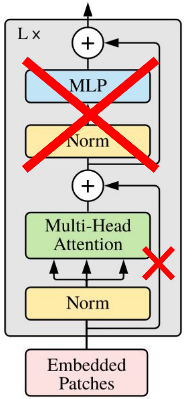
∵ 이 block의 파라미터들이 image-level tasks에 더 최적화 되어있어, dense prediction에 부적절하다고 판단.

2. Skip connection 제거

∵ Skip connection이 previous encoder block의 output을 강조하여, self-attention module output의 중요도를 감소시킨다고 판단.

❑ ∴ *Reduced Architecture* (수식 1~4로부터 개선)

$$\mathbf{Z}^{(L)} = \widetilde{\mathrm{SA}}\left(\mathrm{LN}\left(\mathbf{Z}^{(L-1)}\right)\right), \qquad (16)$$



[1] ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022. p. 696-712.

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

WACV 2025

- **Experiments**

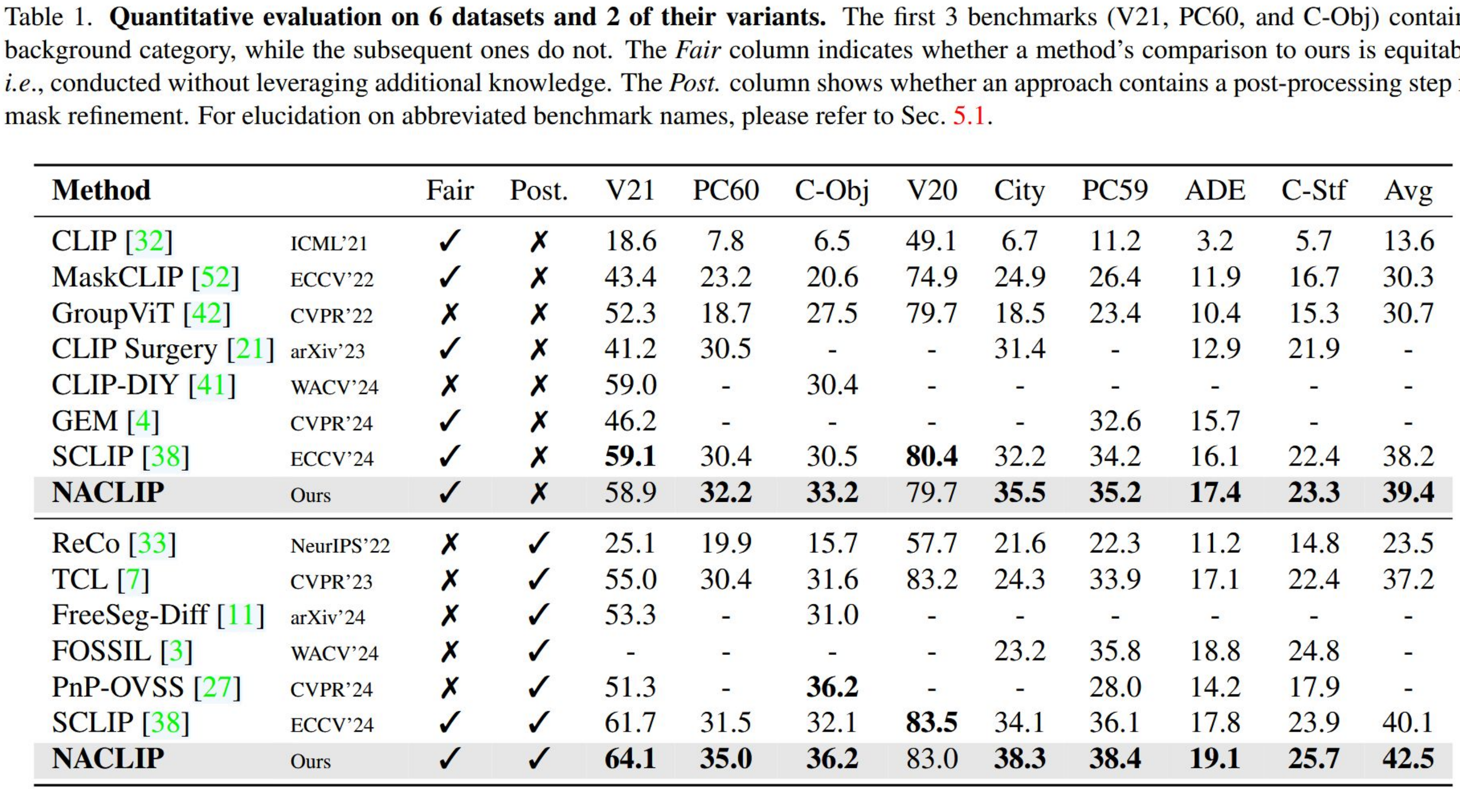


Table 1. **Quantitative evaluation on 6 datasets and 2 of their variants.** The first 3 benchmarks (V21, PC60, and C-Obj) contain a background category, while the subsequent ones do not. The *Fair* column indicates whether a method's comparison to ours is equitable, *i.e.*, conducted without leveraging additional knowledge. The *Post.* column shows whether an approach contains a post-processing step for mask refinement. For elucidation on abbreviated benchmark names, please refer to Sec. 5.1.

Method		Fair	Post.	V21	PC60	C-Obj	V20	City	PC59	ADE	C-Stf	Avg
CLIP [32]	ICML'21	✓	✗	18.6	7.8	6.5	49.1	6.7	11.2	3.2	5.7	13.6
MaskCLIP [52]	ECCV'22	✓	✗	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
GroupViT [42]	CVPR'22	✗	✗	52.3	18.7	27.5	79.7	18.5	23.4	10.4	15.3	30.7
CLIP Surgery [21]	arXiv'23	✓	✗	41.2	30.5	-	-	31.4	-	12.9	21.9	-
CLIP-DIY [41]	WACV'24	✗	✗	59.0	-	30.4	-	-	-	-	-	-
GEM [4]	CVPR'24	✓	✗	46.2	-	-	-	-	32.6	15.7	-	-
SCLIP [38]	ECCV'24	✓	✗	**59.1**	30.4	30.5	**80.4**	32.2	34.2	16.1	22.4	38.2
NACLIP	Ours	✓	✗	58.9	**32.2**	**33.2**	79.7	**35.5**	**35.2**	**17.4**	**23.3**	**39.4**
ReCo [33]	NeurIPS'22	✗	✓	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
TCL [7]	CVPR'23	✗	✓	55.0	30.4	31.6	83.2	24.3	33.9	17.1	22.4	37.2
FreeSeg-Diff [11]	arXiv'24	✗	✓	53.3	-	31.0	-	-	-	-	-	-
FOSSIL [3]	WACV'24	✗	✓	-	-	-	-	23.2	35.8	18.8	24.8	-
PnP-OVSS [27]	CVPR'24	✗	✓	51.3	-	**36.2**	-	-	28.0	14.2	17.9	-
SCLIP [38]	ECCV'24	✓	✓	61.7	31.5	32.1	**83.5**	34.1	36.1	17.8	23.9	40.1
NACLIP	Ours	✓	✓	**64.1**	**35.0**	**36.2**	83.0	**38.3**	**38.4**	**19.1**	**25.7**	**42.5**

전유진

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.

WACV 2025

- **Experiments**

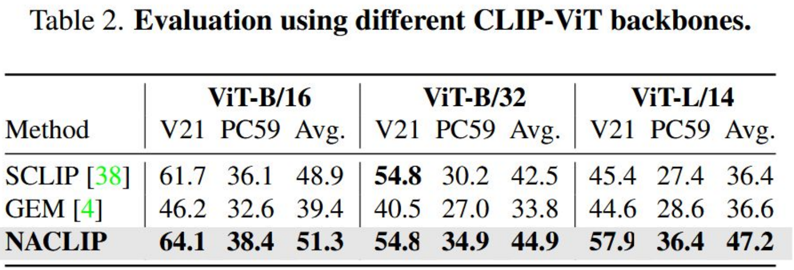


Table 2. **Evaluation using different CLIP-ViT backbones.**

Method	ViT-B/16			ViT-B/32			ViT-L/14		
	V21	PC59	Avg.	V21	PC59	Avg.	V21	PC59	Avg.
SCLIP [38]	61.7	36.1	48.9	**54.8**	30.2	42.5	45.4	27.4	36.4
GEM [4]	46.2	32.6	39.4	40.5	27.0	33.8	44.6	28.6	36.6
NACLIP	**64.1**	**38.4**	**51.3**	**54.8**	**34.9**	**44.9**	**57.9**	**36.4**	**47.2**

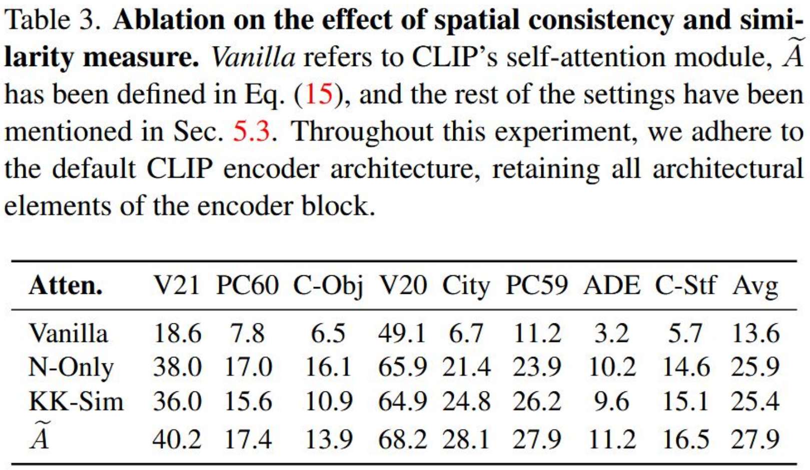


Table 3. **Ablation on the effect of spatial consistency and similarity measure.** *Vanilla* refers to CLIP's self-attention module, $\widetilde{A}$ has been defined in Eq. (15), and the rest of the settings have been mentioned in Sec. 5.3. Throughout this experiment, we adhere to the default CLIP encoder architecture, retaining all architectural elements of the encoder block.

Atten.	V21	PC60	C-Obj	V20	City	PC59	ADE	C-Stf	Avg
Vanilla	18.6	7.8	6.5	49.1	6.7	11.2	3.2	5.7	13.6
N-Only	38.0	17.0	16.1	65.9	21.4	23.9	10.2	14.6	25.9
KK-Sim	36.0	15.6	10.9	64.9	24.8	26.2	9.6	15.1	25.4
$\widetilde{A}$	40.2	17.4	13.9	68.2	28.1	27.9	11.2	16.5	27.9

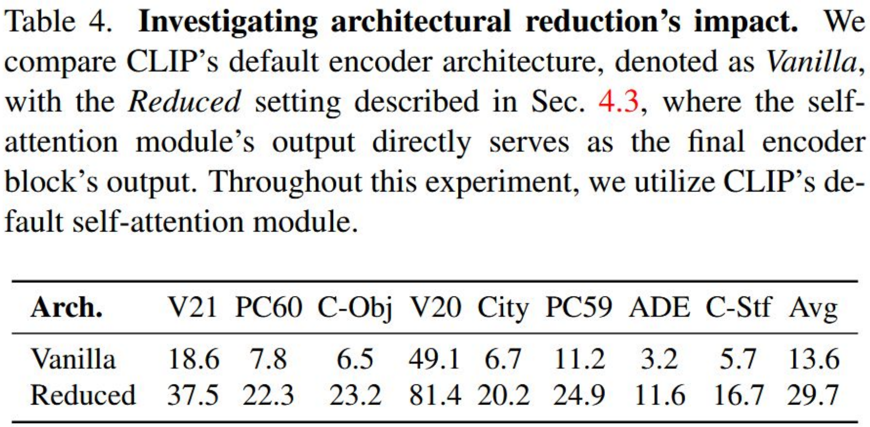


Table 4. **Investigating architectural reduction's impact.** We compare CLIP's default encoder architecture, denoted as *Vanilla*, with the *Reduced* setting described in Sec. 4.3, where the self-attention module's output directly serves as the final encoder block's output. Throughout this experiment, we utilize CLIP's default self-attention module.

Arch.	V21	PC60	C-Obj	V20	City	PC59	ADE	C-Stf	Avg
Vanilla	18.6	7.8	6.5	49.1	6.7	11.2	3.2	5.7	13.6
Reduced	37.5	22.3	23.2	81.4	20.2	24.9	11.6	16.7	29.7

전유진

[HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. p. 5061-5071.]

WACV 2025

- **Experiments**

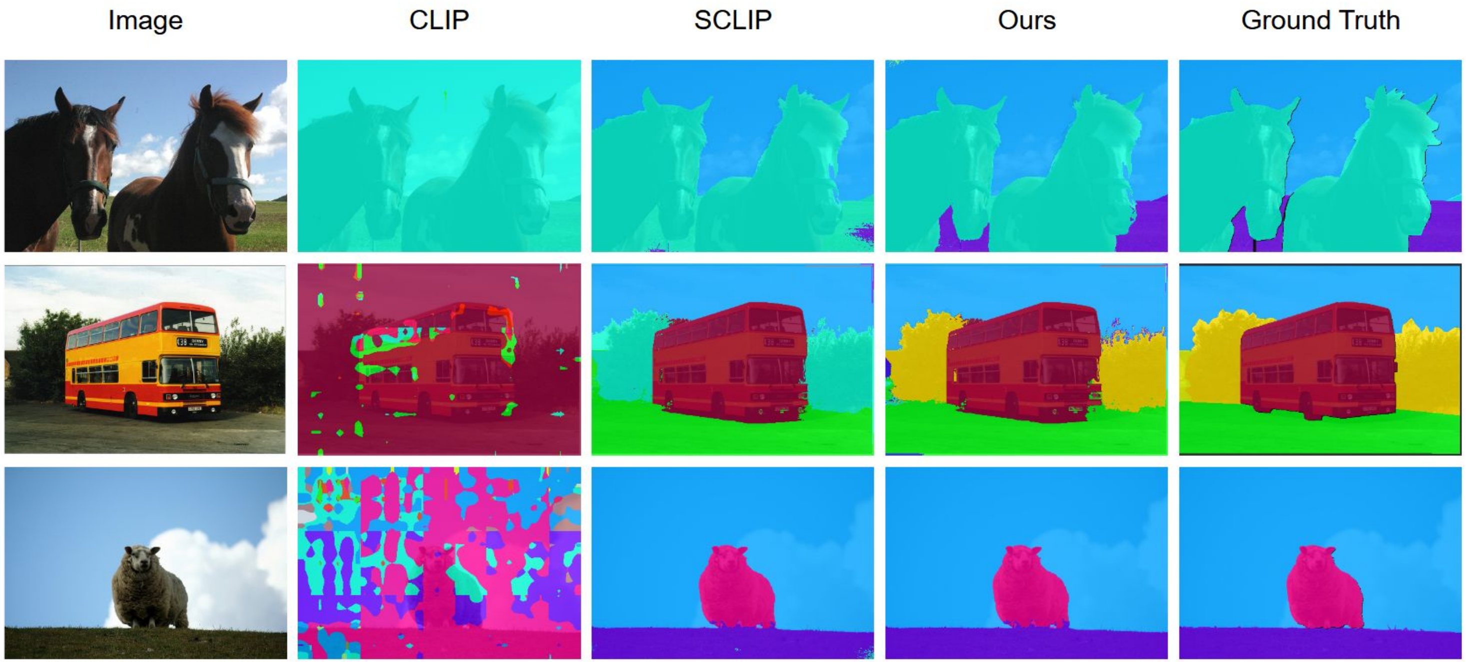


Figure 3. **Qualitative results (segmentation maps) on PASCAL Context (59)** [29] for CLIP [32], SCLIP [38], and our method.

전유진

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, J ion. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

WACV 2025

- **Experiments**

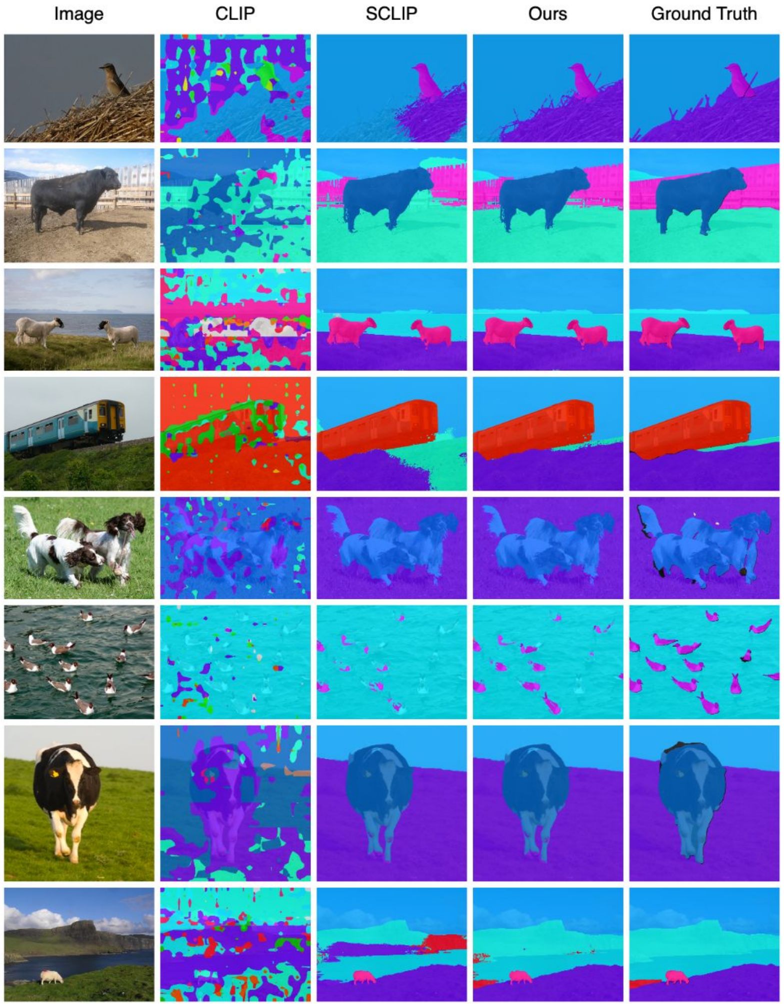


Figure 6. **Additional visual examples (segmentation maps) from PASCAL Context (59)** [29] for CLIP [32], SCLIP [38], and our method.

전유진

HAJIMIRI, Sina; AYED, Ismail Ben; DOLZ, Jose

on Applications of Computer Vision (WACV). IE

ntation. In: *2025 IEEE/CVF Winter Conference*

WACV 2025

- **Experiments**

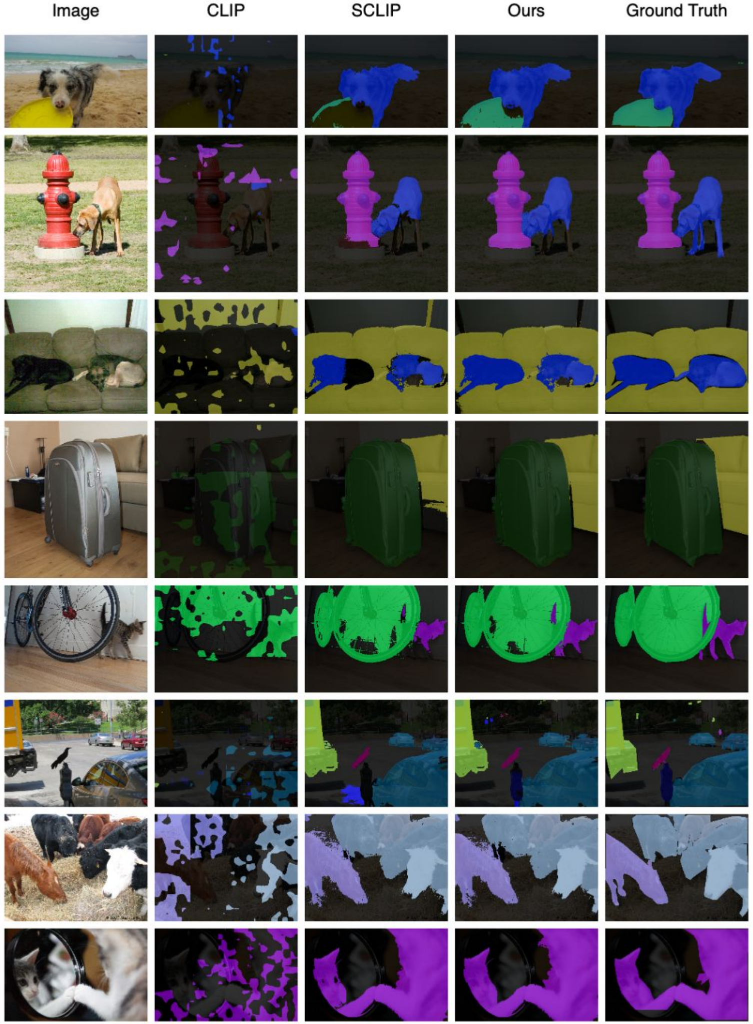


Figure 7. **Additional visual examples (segmentation maps) from COCO-Object** [5, 23] for CLIP [32], SCLIP [38], and our method.

전유진