Seminar

朱顺尧

2026.04.03

Decoupling Template Bias in CLIP: Harnessing Empty Prompts for Enhanced Few-Shot Learning

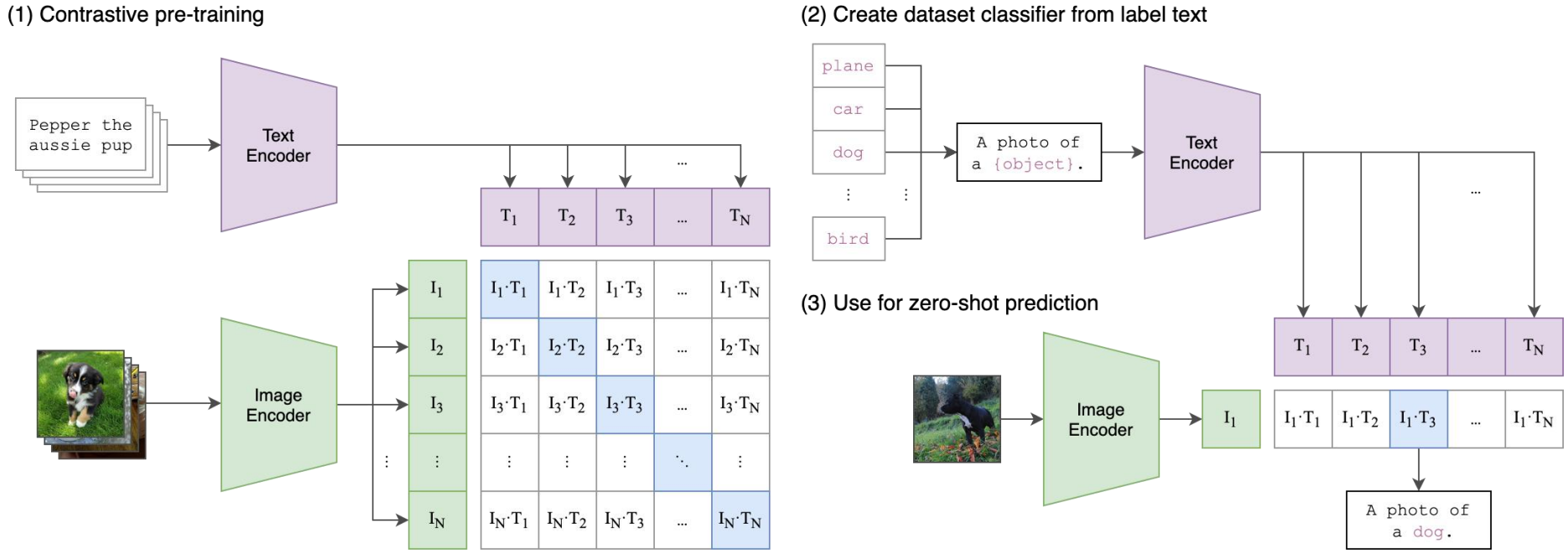
Zhenyu Zhang[1], Guangyao Chen[2*], Yixiong Zou[1*], Zhimeng Huang[2], Yuhua Li[1*]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

AAAI2026

Base Method

CLIP



$$P(y|\boldsymbol{x}, \boldsymbol{p}_k) = \frac{\exp(\langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}_k)\rangle/\gamma)}{\sum_{k'=1}^{K} \exp(\langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}_{k'})\rangle/\gamma)}$$

Motivation

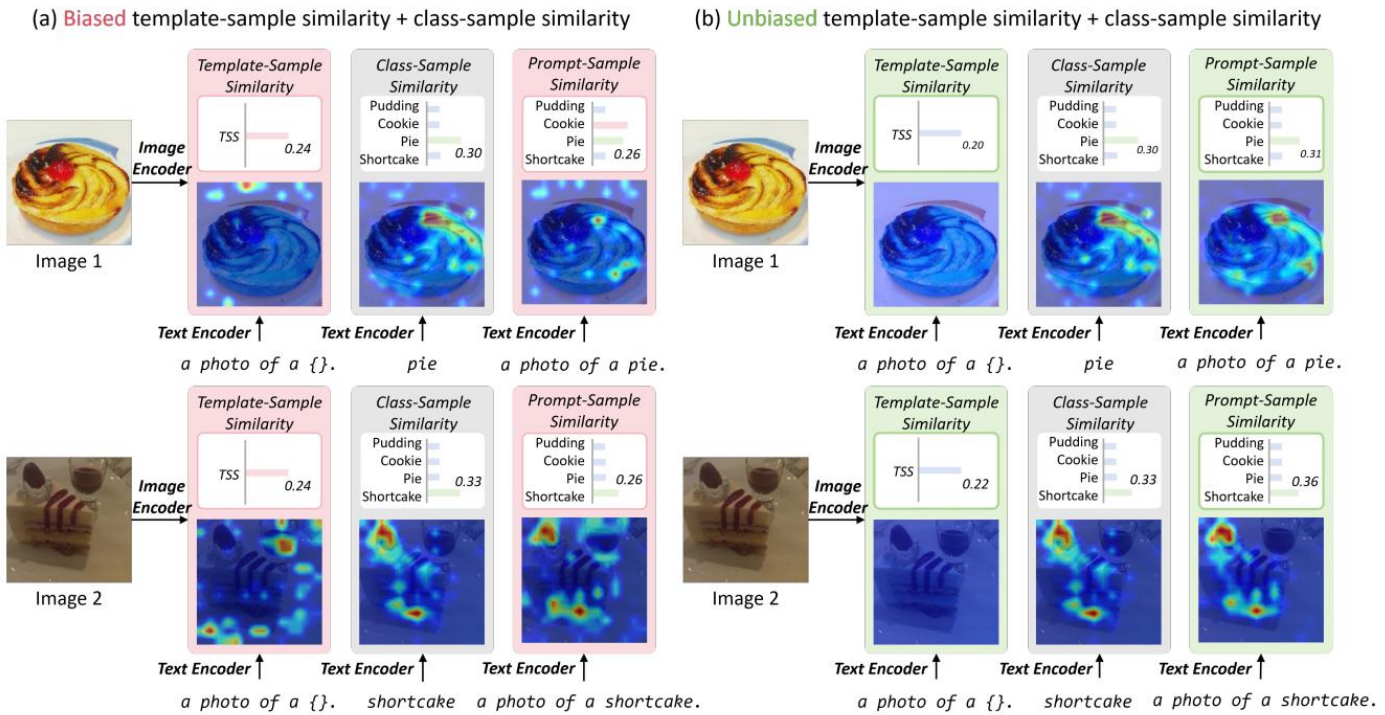


Figure 1: Impact of template bias on CLIP classification. (a) A biased prompt ("a photo of a ") introduces extra template–sample similarity that skews attention toward irrelevant regions and inflates incorrect class scores—even when the class name matches the object. (b) An unbiased prompt removes template bias, so the model focuses on genuinely discriminative features and the model's prediction relies solely on class–sample similarity and correct prompt–sample similarity. .

$$\mathrm{TSS}(s) = \cos\big(\theta_t(t_0), \theta_v(s)\big),$$

$$\mathrm{CSS}(c, s) = \cos\big(\theta_t(c), \theta_v(s)\big),$$

$$\mathrm{PSS}(c, s) = \cos\big(\theta_t(p_c), \theta_v(s)\big).$$

	ImagNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Avg.
Only class	64.08	60.81	20.97	**46.38**	63.67	83.94	81.92	64.67	87.22	**45.80**	63.83	62.12
w/ Template	**66.68**	**62.49**	**23.25**	42.38	**65.40**	**85.27**	**88.36**	**67.35**	**93.38**	44.32	**65.15**	64.00
Misclassification Ratio	4.82	6.49	4.32	9.79	5.19	2.39	1.47	4.1	1.29	8.21	6.18	4.93

Table 1: Comparison of classification performance using bare class names versus full templates. We define the *misclassification ratio* as **(samples correctly classified using only class names but misclassified after introducing templates) / (total samples)**. Although templates improve overall accuracy, they also induce errors on a subset of samples that were previously classified correctly.

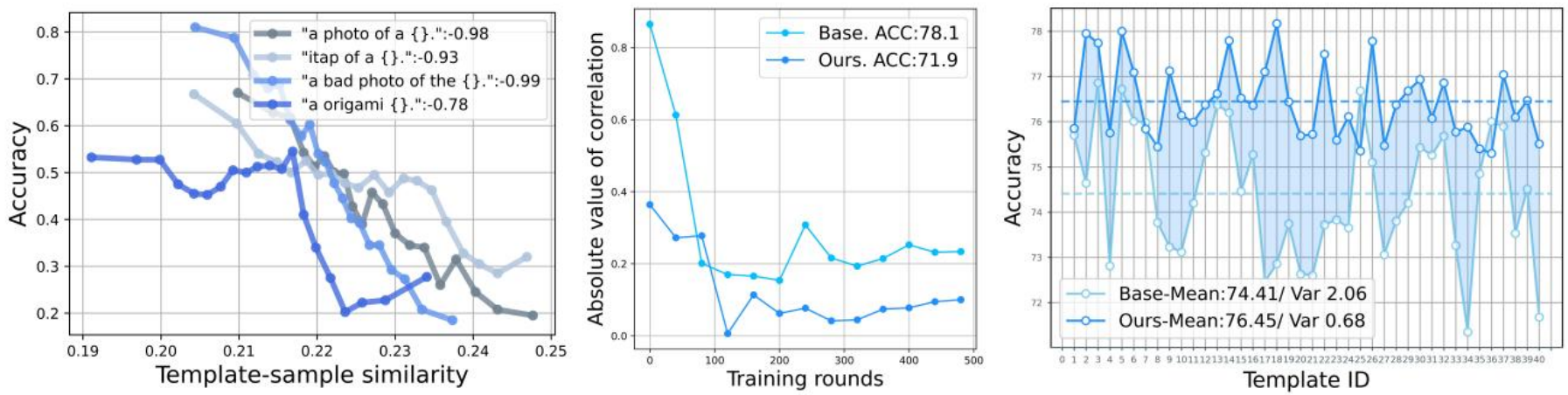


Figure 2: (Left) Correlation between Template-Sample Similarity and classification accuracy on EuroSAT (correlation coefficients is included in the legend). (Middle) The evolution of the absolute value of correlation coefficients between classification accuracy and template-sample similarity over the course of 1-shot training on EuroSAT. (Right) The effect of different templates on overall performance, shown both before and after applying template correction on EuroSAT.

Method

Empty Prompts Generation

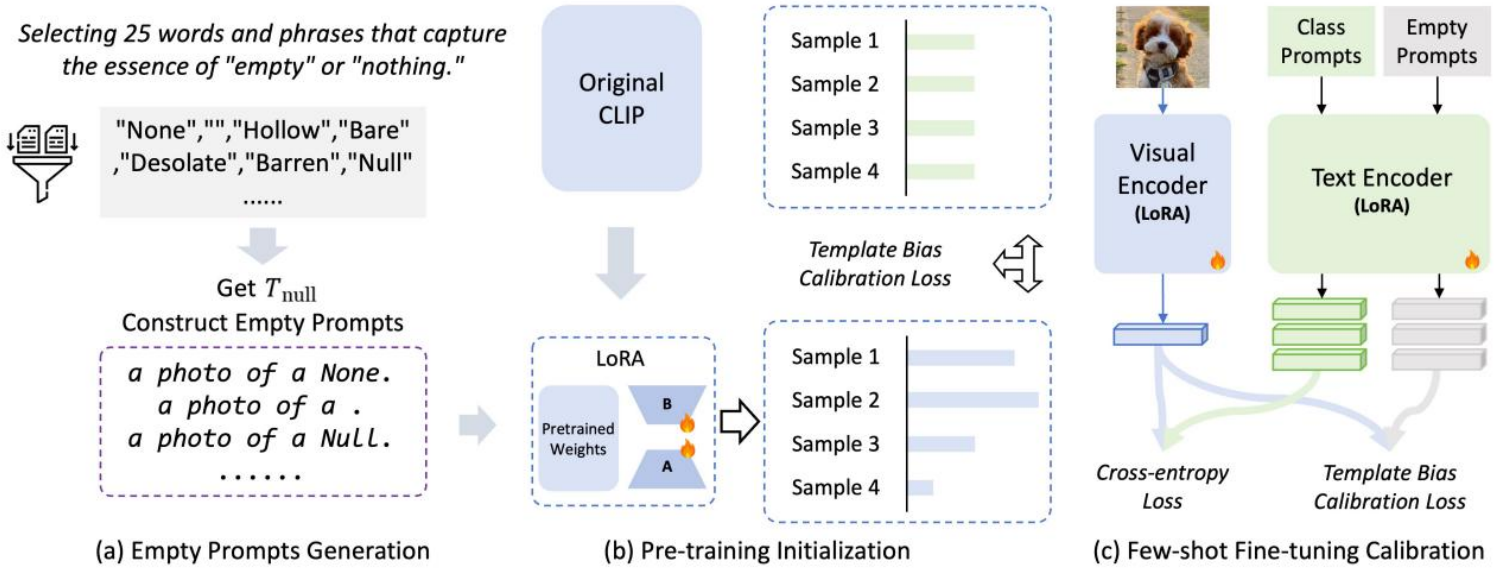


Figure 3: The overall framework for template correction involves three main stages. (a) **Empty Prompts Generation**: A diverse set of empty prompts is manually curated to help identify potential template-induced biases. (b) **Pretraining Initialization**: These empty prompts are used to detect and correct biases within the CLIP model during the pretraining phase. (c) **Few-shot Fine-tuning Calibration**: Finally, the model undergoes fine-tuning with few-shot samples to calibrate its performance and improve classification accuracy.

We employ the following 25 terms to represent "empty": "None", " ", "Vacant", "BlankVoid", "Hollow", "Bare", "Desolate", "Barren", "Null", "Naked", "Devoid", "Vacuous", "Unoccupied", "Sparse", "Clean", "Clear", "Abandoned", "Forsaken", "Deserted", "Uninhabited", "Unused", "Vast", "Sterile", "Unfilled", "Unpopulated". For example, using the template "a photo of a ", the corresponding empty prompts are:

```
a photo of a None.
a photo of a .
a photo of a Vacant.
a photo of a BlankVoid.
a photo of a Hollow.
a photo of a Bare.
a photo of a Desolate.
a photo of a Barren.
a photo of a Null.
a photo of a Naked.
a photo of a Devoid.
a photo of a Vacuous.
a photo of a Unoccupied.
a photo of a Sparse.
a photo of a Clean.
a photo of a Clear.
a photo of a Abandoned.
a photo of a Forsaken.
a photo of a Deserted.
a photo of a Uninhabited.
a photo of a Unused.
a photo of a Vast.
a photo of a Sterile.
a photo of a Unfilled.
a photo of a Unpopulated.
```

Method

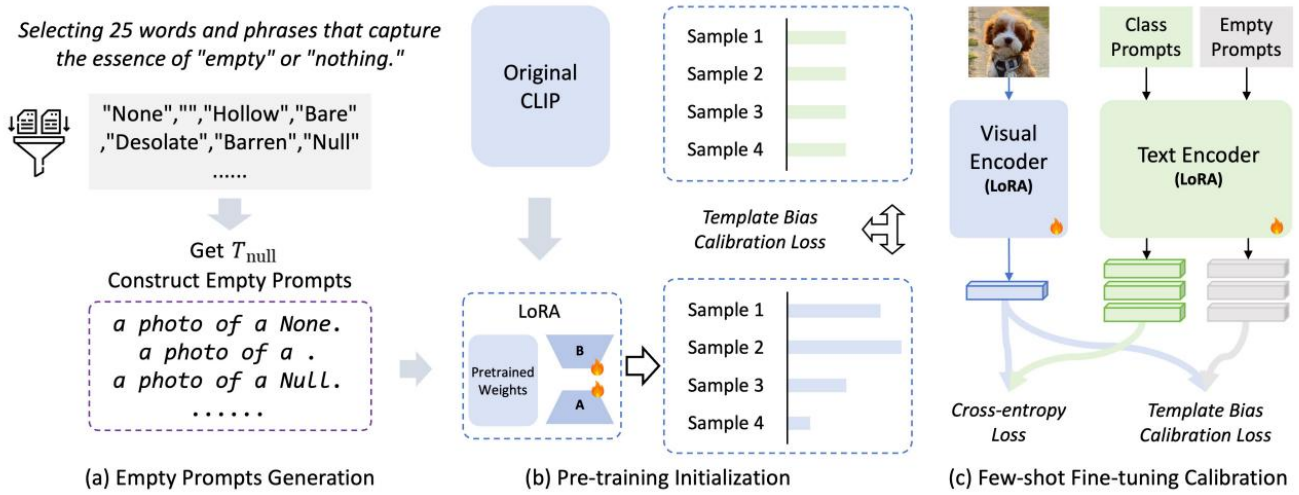


Figure 3: The overall framework for template correction involves three main stages. (a) **Empty Prompts Generation**: A diverse set of empty prompts is manually curated to help identify potential template-induced biases. (b) **Pretraining Initialization**: These empty prompts are used to detect and correct biases within the CLIP model during the pretraining phase. (c) **Few-shot Fine-tuning Calibration**: Finally, the model undergoes fine-tuning with few-shot samples to calibrate its performance and improve classification accuracy.

$$\mathbb{E}_{\mathbf{t}_i^E}\left[P\left(y \mid \mathbf{t}_i^E, \mathcal{M}; \forall y \in \mathcal{Y}_x\right)\right] = \frac{1}{|\mathcal{Y}_x|}$$

$$L_{tb} = -\frac{1}{|T_{null}|}\sum_{i=1}^{|T_{null}|}\sum_{k=1}^{|\mathcal{Y}_x|}\frac{1}{|\mathcal{Y}_x|}\ln p_{i,k}^{E}$$

Method

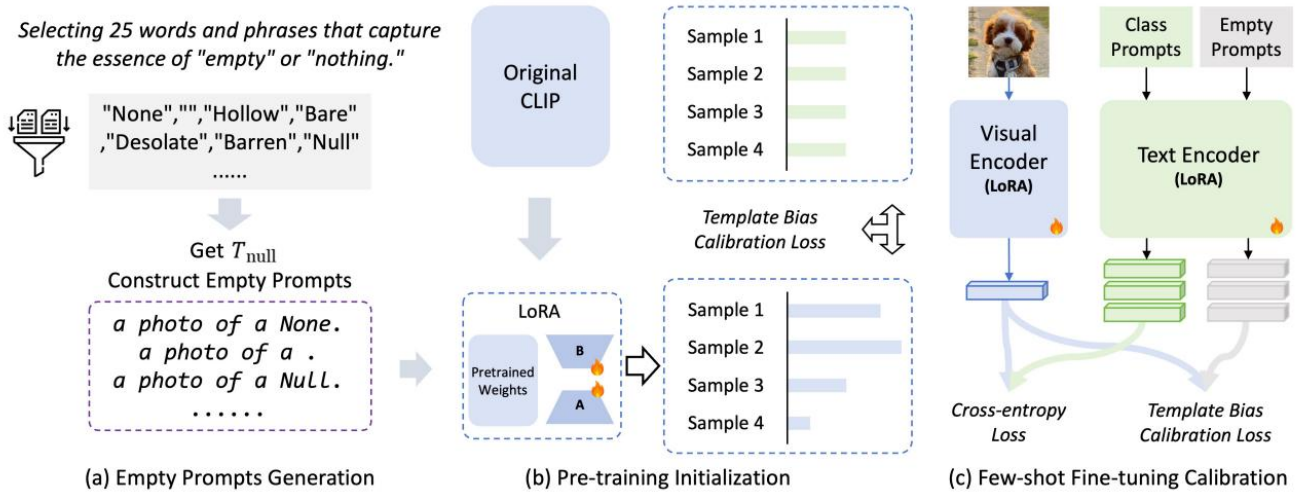


Figure 3: The overall framework for template correction involves three main stages. (a) **Empty Prompts Generation**: A diverse set of empty prompts is manually curated to help identify potential template-induced biases. (b) **Pretraining Initialization**: These empty prompts are used to detect and correct biases within the CLIP model during the pretraining phase. (c) **Few-shot Fine-tuning Calibration**: Finally, the model undergoes fine-tuning with few-shot samples to calibrate its performance and improve classification accuracy.

Pre-training Initializaiton (Unsupervised)

$$L_{tb} = -\frac{1}{|T_{null}|} \sum_{i=1}^{|T_{null}|} \sum_{k=1}^{|\mathcal{Y}_x|} \frac{1}{|\mathcal{Y}_x|} \ln p_{i,k}^{E}$$

Few-shot Fine-tuning Calibration

$$L_{fine} = L_{ce} + \alpha L_{tb}$$

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \ln p_{i,k}$$

Results

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
	CoOp(4)	68.0	67.3	26.2	50.9	67.1	82.6	90.3	72.7	93.2	50.1	70.7	67.2
	CoOp(16)	65.7	67.0	20.8	56.4	67.5	84.3	90.2	78.3	92.5	50.1	71.2	67.6
	CoCoOp	69.4	68.7	28.1	55.4	67.6	84.9	91.9	73.4	94.1	52.6	70.4	68.8
	TIP-Adapter-F	69.4	67.2	28.8	67.8	67.1	85.8	90.6	**83.8**	94.0	51.6	73.4	70.9
	CLIP-Adapter	67.9	65.4	25.2	49.3	65.7	86.1	89.0	71.3	92.0	44.2	66.9	65.7
1	PLOT++	66.5	66.8	28.6	65.4	68.8	86.2	91.9	80.5	94.3	54.6	74.3	70.7
	KgCoOp	68.9	68.4	26.8	61.9	66.7	**86.4**	92.1	74.7	94.2	52.7	72.8	69.6
	TaskRes	69.6	68.1	31.3	65.4	68.8	84.6	90.2	81.7	93.6	53.8	71.7	70.8
	MaPLe	69.7	69.3	28.1	29.1	67.6	85.4	91.4	74.9	93.6	50.0	71.1	66.4
	ProGrad	67.0	67.0	28.8	57.0	68.2	84.9	91.4	80.9	93.5	52.8	73.3	69.5
	LoRA-CLIP	70.4	70.4	30.2	72.3	70.1	84.3	92.3	83.2	93.7	54.3	**76.3**	72.5
	OURS	**70.5**	**70.5**	**32.2**	**78.4**	**70.6**	85.7	**92.9**	83.3	**94.8**	**54.5**	76.0	**73.6**
	CoOp(4)	68.7	68.0	28.1	66.2	70.5	82.6	89.9	80.9	93.0	53.7	73.5	70.5
	CoOp(16)	67.0	67.0	25.9	65.1	70.4	84.4	89.9	88.0	93.1	54.1	74.1	70.8
	CoCoOp	70.1	69.4	29.3	61.8	68.4	85.9	91.9	77.8	94.4	52.3	73.4	70.4
	TIP-Adapter-F	70.0	68.6	32.8	73.2	70.8	86.0	91.6	90.1	93.9	57.8	76.2	73.7
	CLIP-Adapter	68.2	67.2	27.0	51.2	66.6	86.2	89.7	71.7	93.4	45.4	68.4	66.8
2	PLOT++	68.3	68.1	31.1	76.8	73.2	86.3	92.3	89.8	94.7	56.7	76.8	74.0
	KgCoOp	69.6	69.6	28.0	69.2	68.2	**86.6**	**92.3**	79.8	94.5	55.3	74.6	71.6
	TaskRes	70.2	70.5	32.7	70.2	72.1	85.6	90.7	84.4	94.3	55.6	75.2	72.9
	MaPLe	70.0	70.7	29.5	59.4	68.5	86.5	91.8	79.8	94.9	50.6	74.0	70.5
	ProGrad	69.1	69.0	31.1	66.3	72.4	84.8	91.5	87.5	93.6	56.0	75.6	72.4
	CLIP-LoRA	70.8	71.3	33.2	82.7	73.2	83.2	91.3	89.8	94.6	59.9	**80.0**	75.5
	Ours	**71.0**	**71.7**	**34.9**	**83.3**	**73.5**	86.2	91.9	**90.4**	**95.1**	**61.4**	79.8	**76.3**
	CoOp (4)	69.7	70.6	29.7	65.8	73.4	83.5	92.3	86.6	94.5	58.5	78.1	73.0
	CoOp (16)	68.8	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.0
	CoCoOp	70.6	70.4	30.6	61.7	69.5	86.3	92.7	81.5	94.8	55.7	75.3	71.7
	TIP-Adapter-F	70.7	70.8	35.7	76.8	74.1	86.5	91.9	92.1	94.8	59.8	78.1	75.6
	CLIP-Adapter	68.6	68.0	27.9	51.2	67.5	86.5	90.8	73.1	94.0	46.1	70.6	67.7
4	PLOT++	70.4	71.7	35.3	83.2	76.3	86.5	92.6	92.9	95.1	62.4	79.8	76.9
	KgCoOp	69.9	71.5	32.2	71.8	69.5	86.9	92.6	87.0	95.0	58.7	77.6	73.9
	TaskRes	71.0	72.7	33.4	74.2	76.0	86.0	91.9	85.0	95.0	60.1	76.2	74.7
	MaPLe	70.6	71.4	30.1	69.9	70.1	86.7	93.3	84.9	95.0	59.0	77.1	73.5
	ProGrad	70.2	71.7	34.1	69.6	75.0	85.4	92.1	91.1	94.4	59.7	77.9	74.7
	LoRA-CLIP	71.4	72.8	37.9	84.9	77.4	82.7	91.0	93.7	95.2	63.8	81.1	77.4
	OURS	**71.5**	**74.0**	**40.2**	**87.6**	**77.7**	**87.1**	**93.7**	**94.8**	**95.7**	**65.6**	**82.2**	**79.1**
	CoOp(4)	70.8	72.4	37.0	74.7	76.8	83.3	92.1	95.0	94.7	63.7	79.8	76.4
	CoOp(16)	70.6	71.9	38.5	77.1	79.0	82.7	91.3	94.9	94.5	64.8	80.0	76.8
	CoCoOp	70.8	71.5	32.4	69.1	70.4	87.0	93.3	86.3	94.9	60.1	75.9	73.8
	TIP-Adapter-F	71.7	73.5	39.5	81.3	78.3	86.9	91.8	94.3	95.2	66.7	82.0	78.3
	CLIP-Adapter	69.1	71.7	30.5	61.6	70.7	86.9	91.9	83.3	94.5	50.5	76.2	71.5
8	PLOT++	71.3	73.9	41.4	88.4	81.3	86.6	93.0	95.4	95.5	66.5	82.8	79.6
	KgCoOp	70.2	72.6	34.8	73.9	72.8	87.0	93.0	91.5	95.1	65.6	80.0	76.0
	TaskRes	72.3	74.6	40.3	77.5	79.6	86.4	92.0	96.0	95.3	66.7	81.6	78.4
	MaPLe	71.3	73.2	33.8	82.8	71.3	87.2	93.1	90.5	95.1	63.0	79.5	76.4
	ProGrad	71.3	73.0	37.7	77.8	78.7	86.1	92.2	95.0	94.8	63.9	80.5	77.4
	CLIP-LoRA	72.3	74.7	45.7	89.7	82.1	83.1	91.7	96.3	95.6	67.5	84.1	80.3
	OURS	**72.3**	**75.5**	**49.7**	**91.2**	**82.5**	**87.3**	**94.2**	**96.9**	**95.9**	**69.2**	**85.0**	**81.8**
	CoOp(4)	71.5	74.6	40.1	83.5	79.1	85.1	92.4	96.4	95.5	69.2	81.9	79.0
	CoOp(16)	71.9	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.0
	CoCoOp	71.1	72.6	33.3	73.6	72.3	87.4	93.4	89.1	95.1	63.7	77.2	75.4
	TIP-Adapter-F	73.4	76.0	44.6	85.9	82.3	86.8	92.6	96.2	95.7	70.8	83.9	80.7
	CLIP-Adapter	69.8	74.2	34.2	71.4	74.0	87.1	92.3	92.9	94.9	59.4	80.2	75.5
16	PLOT++	72.6	76.0	46.7	92.0	84.6	87.1	93.6	97.6	96.0	71.4	85.3	82.1
	KgCoOp	70.4	73.3	36.5	76.2	74.8	87.2	93.2	93.4	95.2	68.7	81.7	77.3
	TaskRes	73.0	76.1	44.9	82.7	83.5	86.9	92.4	97.5	95.8	71.5	84.0	80.8
	MaPLe	71.9	74.5	36.8	87.5	74.3	87.4	93.2	94.2	95.4	68.4	81.4	78.6
	ProGrad	72.1	75.1	43.0	83.6	82.9	85.8	92.8	96.6	95.9	68.8	82.7	79.9
	CLIP-LoRA	73.6	76.1	54.7	92.1	86.3	84.2	92.4	98.0	96.4	72.0	**86.7**	83.0
	Ours	**73.6**	**76.4**	**57.4**	**92.4**	**86.3**	**87.6**	**94.6**	**98.2**	**96.5**	**72.8**	86.5	**83.8**

Table 2: Detailed results for 11 datasets with the ViT-B/16 as visual backbone.Top-1 accuracy averaged over 3 random seeds is reported. Highest value is high lighted inbold.

Ablation

Pretrain	-	-	-	✓	✓	✓
Multiple	-	-	✓	-	✓	✓
$+L_{tb}$	-	✓	✓	-	-	✓
1-shot	72.3	73.9	76.2	73.3	74.2	78.4
4-shot	84.9	85.8	86.9	86.4	87.9	87.6

Table 3: Ablating main components of our model on EuroSAT.

Method	ImageNet	SUN	Aircraft	EuroSAT	DTD	Average
Baseline	70.4	70.4	30.2	72.3	54.3	59.5
Pull Closer	69.5	69.3	30.0	71.6	51.9	58.5
Push Away	68.4	65.8	7.9	62.9	48.4	50.7
Ours	70.5	70.5	32.2	78.3	54.5	61.2

Table 4: Evaluation of three TSS modification strategies.

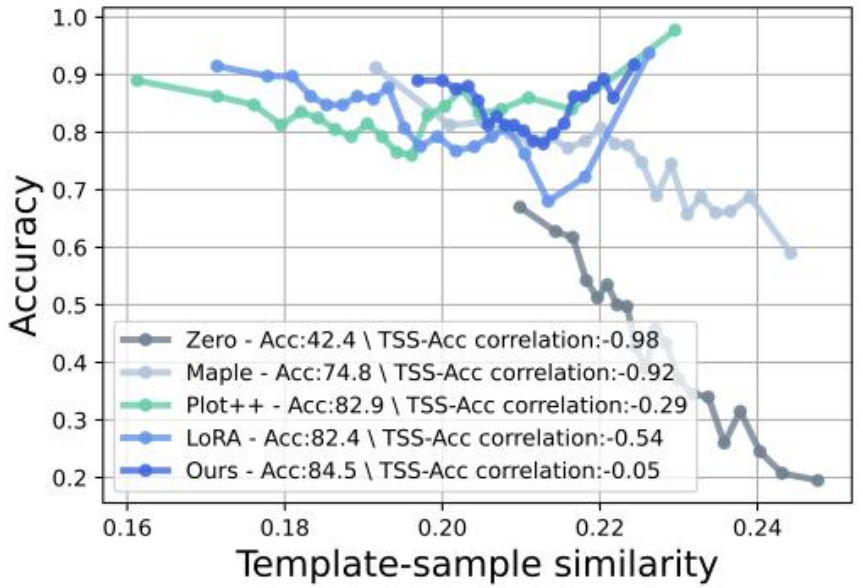


Figure 4: The relationship between template-sample similarity and classification accuracy under different training methods. Result for 4-shot finetuning on EuroSAT, seed 1.

Pull Closer：reduces the distance between each sample and the “empty template” (increasing TSS)

Push Away：increases the distance (decreasing TSS)

Ours：enforcing a consistent TSS

Making Training-Free Diffusion Segmentors Scale with the Generative Power

Benyuan Meng[1,2]
mengbenyuan@iie.ac.cn

Qianqian Xu[3*]
xuqianqian@ict.ac.cn

Zitai Wang[3]
wangzitai@ict.ac.cn

Xiaochun Cao[4]
caoxiaochun@mail.sysu.edu.cn

Longtao Huang[5]
kaiyang.hlt@alibaba-inc.com

Qingming Huang[3,6,7*]
qmhuang@ucas.ac.cn

CVPR2026

Motivation

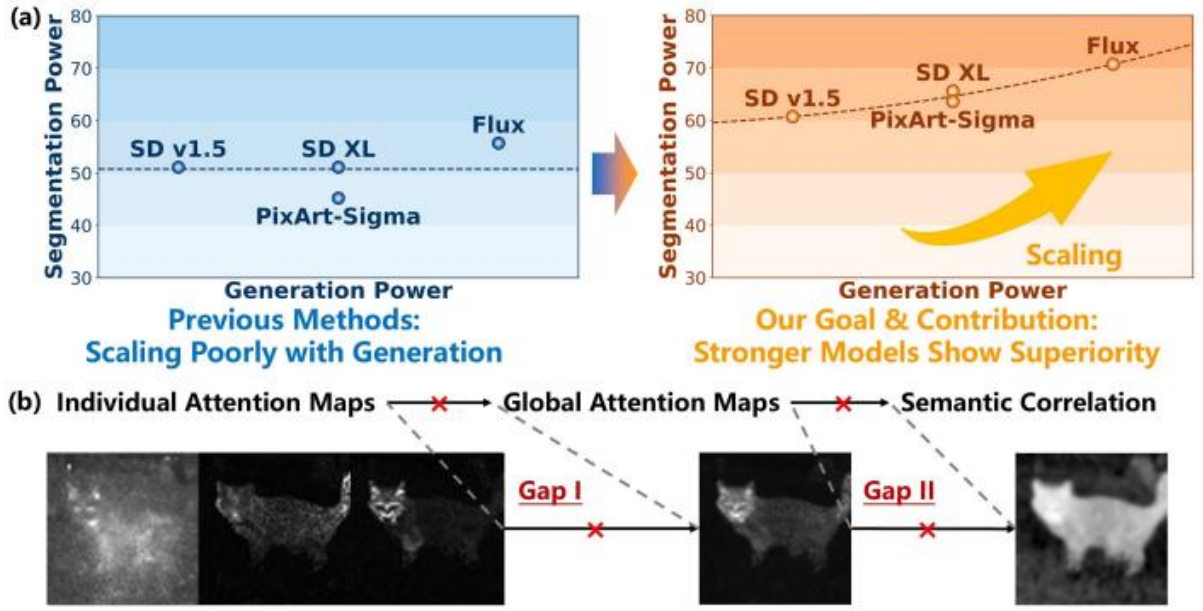


Figure 1. (a) Previous training-free diffusion segmentors scale poorly with the generative power of diffusion models, which inspires our study to enable such scaling. (b) We have identified two gaps from individual cross-attention maps to semantic correlation, which have been preventing the aforementioned scaling.

Method

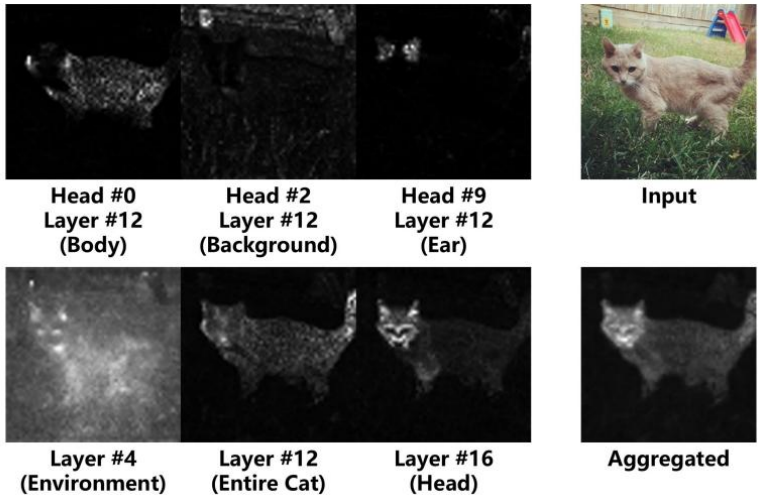
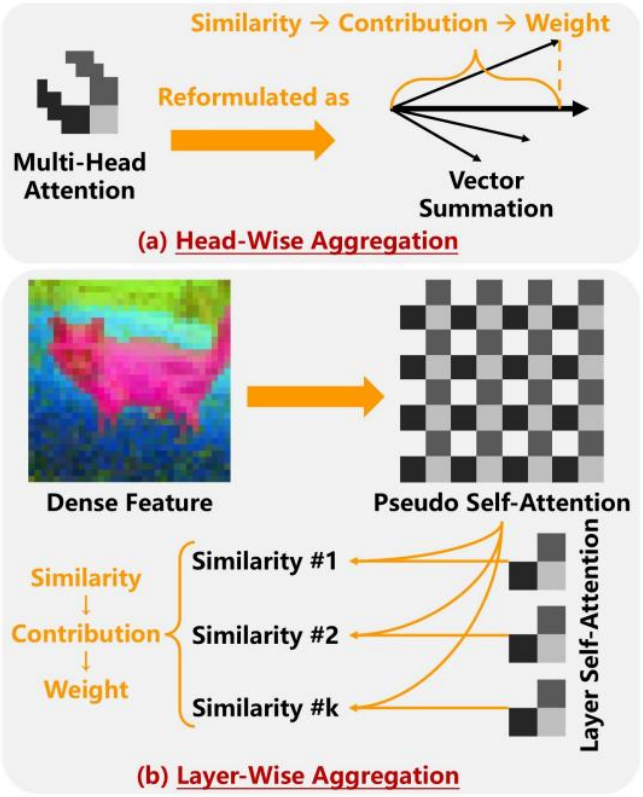
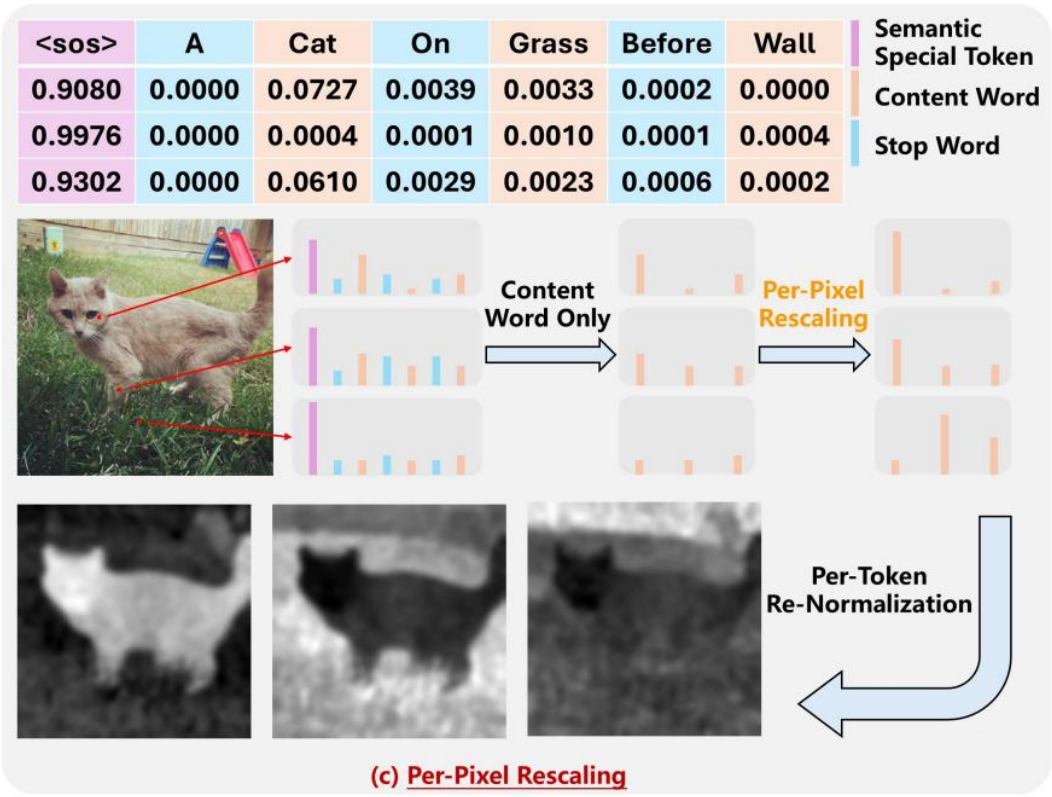


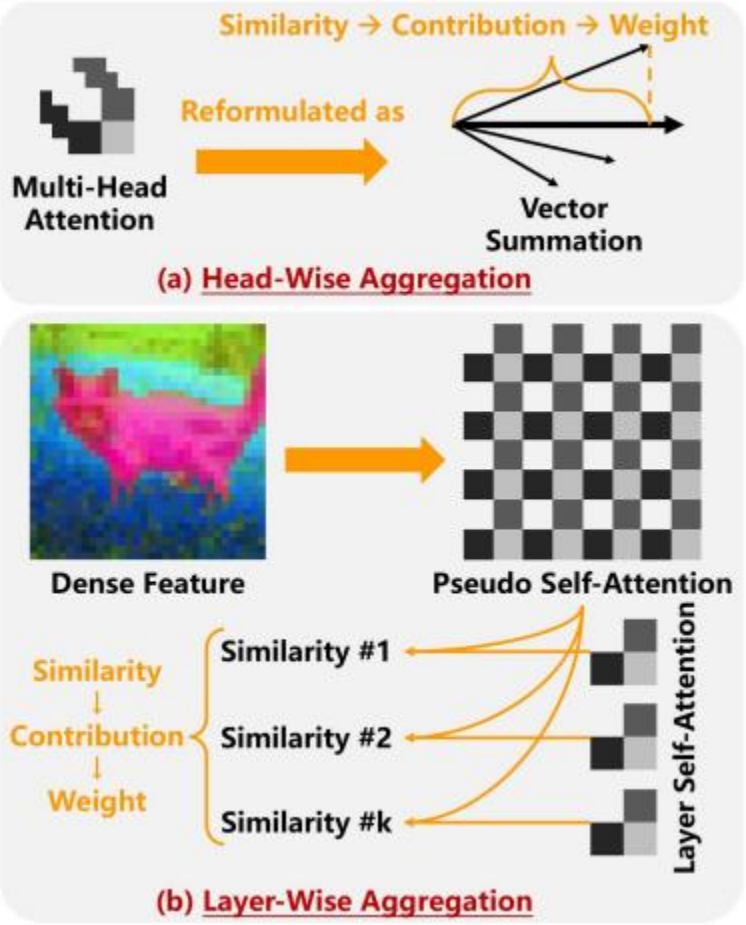
Figure 2. Attention maps in different heads and layers show a certain collaboration pattern, each focusing on distinct aspects of the image.





Method

Addressing Gap 1



$$Output = Concat(\mathbf{A}_1\mathbf{V}_1, \cdots, \mathbf{A}_h\mathbf{V}_h)\mathbf{W}^O.$$

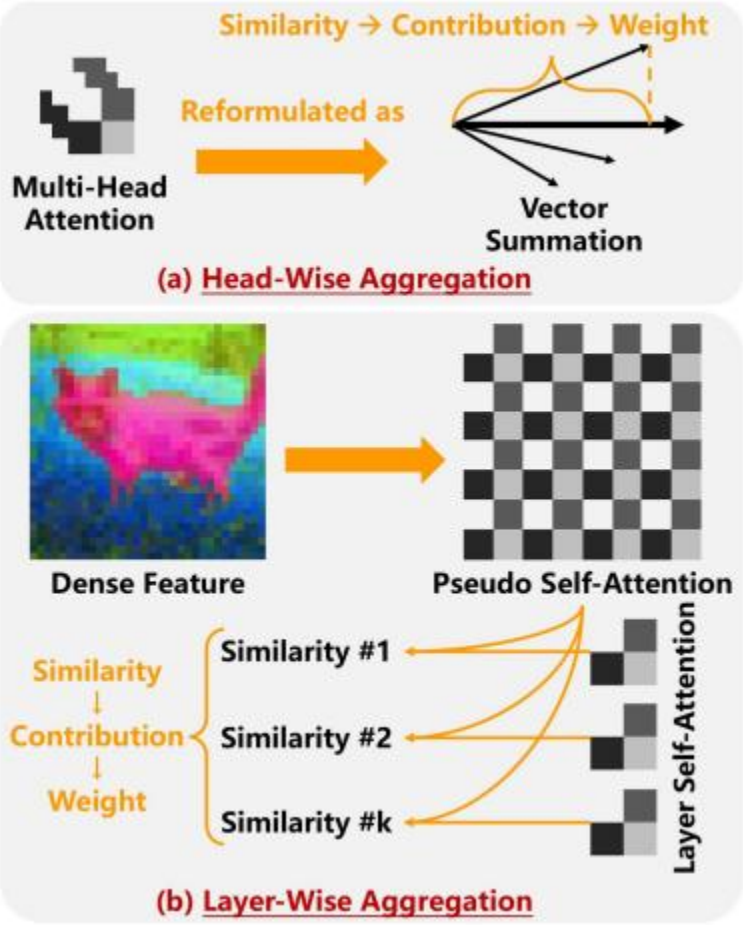
$$Output = \mathbf{A}_1\mathbf{V}_1\mathbf{W}_1^O + \cdots + \mathbf{A}_h\mathbf{V}_h\mathbf{W}_h^O.$$

$$w_{mn} = (\mathbf{A}_n\mathbf{V}_n\mathbf{W}_n^O)_m^T Output_m.$$

$$\mathbf{A}_m = \sum_{n=1}^{h} w'_{mn}\mathbf{A}_{mn}. \qquad w'_{mn} = \frac{w_{mn}}{\sum_{j=1}^{h} w_{mj}}$$

Method

Addressing Gap 1



$$\mathbf{A}_p = Softmax(\frac{feat(feat)^T}{\sqrt{d}}) \in \mathbb{R}^{hw \times hw}.$$

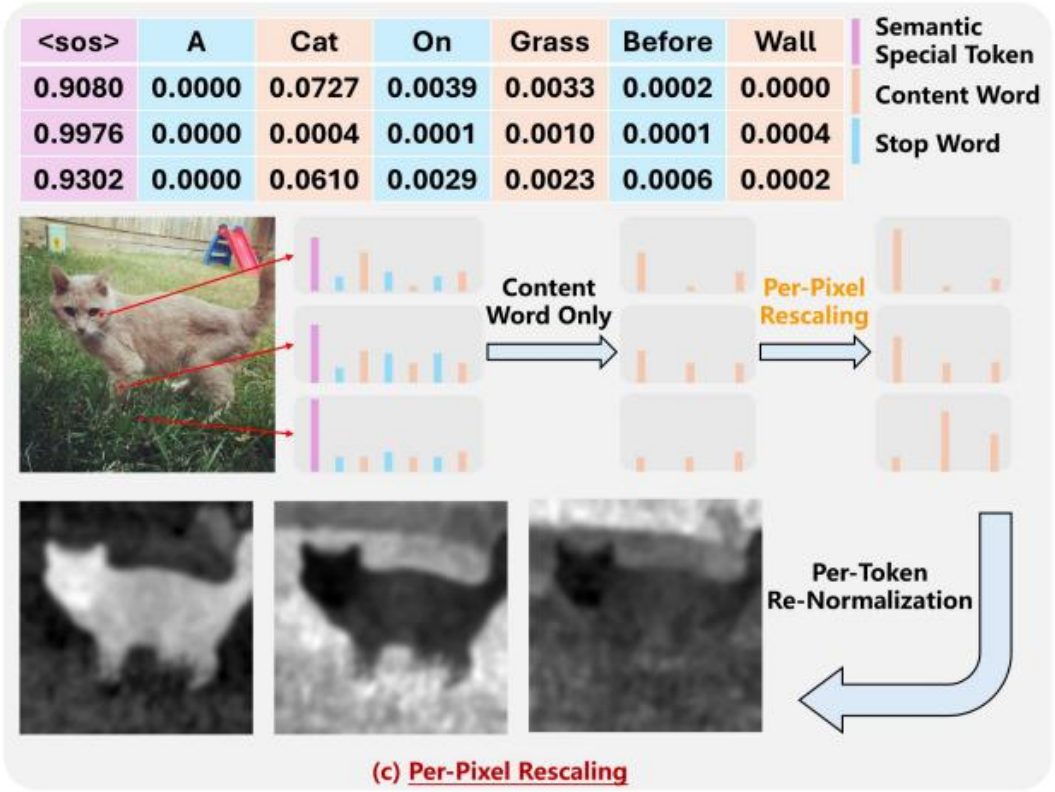
$$w_m = (A'_p)^T (A^m_{self})' \in \mathbb{R},$$

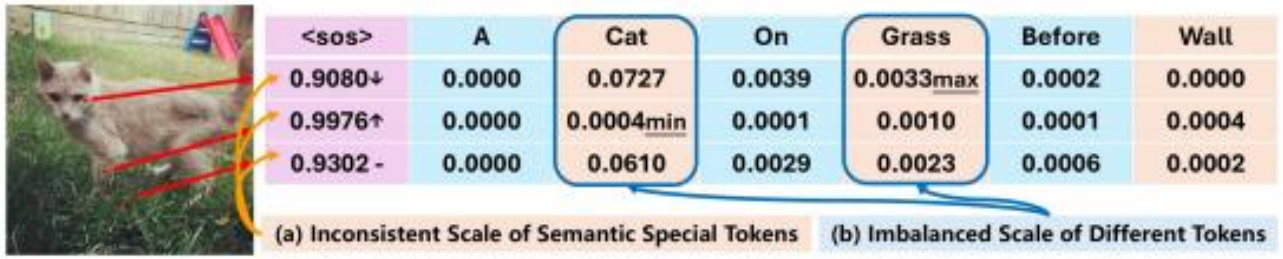
$$w'_m = \frac{w_m}{\sum_{j=1}^{k} w_j}.$$

$$\mathbf{A} = \sum_{m=1}^{k} w'_m \mathbf{A}_m.$$

Method

Addressing Gap 2





$$\mathbf{A}'(i,\boldsymbol{q}) = \frac{\mathbf{A}(i,\boldsymbol{q})}{\sum_{j=1}^{l}\mathbf{A}(i,\boldsymbol{q}(j))}.$$

$$\mathbf{A}''(i,j) = \frac{\mathbf{A}'(i,j) - \min\limits_{i'}(\mathbf{A}'(i',j))}{\max\limits_{i'}(\mathbf{A}'(i',j)) - \min\limits_{i'}(\mathbf{A}'(i',j))}.$$

$$y_i = \underset{j}{argmax}\,\mathbf{A}''(i,j).$$

Results

Table 1. Experimental results on standard semantic segmentation benchmarks, evaluated using mIoU↑. The best and runner-up are **bold** and underlined. [1]Reproduced by us because the original setting is different, details in Section A.

Type	Method	VOC	Context	COCO-Object	Cityscapes	ADE20K
Non-DM	MaskCLIP	38.8	23.6	20.6	10.0	9.8
	ReCO	25.1	19.9	15.7	19.3	11.2
Pre-Trained DM	DiffSegmentor	60.1	27.5	37.9	-	-
	MaskDiffusion	29.9	-	-	17.1	-
	FTTM[1]	48.9	30.0	34.6	12.3	20.3
Vanilla	SD v1.5	44.3	32.3	32.3	11.8	18.0
	SD XL	51.1	35.7	37.2	16.1	18.6
	Pixart-Sigma	45.2	37.0	33.4	22.5	19.1
	Flux	55.7	48.4	43.3	25.6	24.5
Baseline	SD v1.5	51.1	35.4	36.9	18.4	21.0
Ours	SD v1.5	60.7	40.4	39.2	16.1	22.0
	SD XL	65.6	42.3	44.3	21.2	23.2
	Pixart-Sigma	63.6	43.2	39.8	22.6	23.8
	Flux	**70.7**	**51.1**	**48.1**	**27.1**	**29.3**

Ablation

Head	Layer	Rescaling	SD v1.5	SD XL
Vanilla	Vanilla	Vanilla	44.3	51.1
Vanilla	**Manual**	Vanilla	51.1	-
Ours	Vanilla	Vanilla	44.8	56.1
Vanilla	**Ours**	Vanilla	52.1	51.3
Vanilla	Vanilla	**Ours**	52.6	51.4
Ours	**Ours**	**Ours**	**60.7**	**65.6**

Table 5. Ablation results to compare different similarity metrics for head- and layer-wise aggregation.

Model	Head Metric	Layer Metric	mIoU (Context)
SD v1.5	dot-product	dot-product	40.4
	l2-norm	dot-product	**42.0**
	cosine	dot-product	**42.0**
	dot-product	mse	33.9
	dot-product	iou-like	39.2
SD XL	dot-product	dot-product	**42.3**
	l2-norm	dot-product	40.3
	cosine	dot-product	40.3
	dot-product	mse	42.1
	dot-product	iou-like	41.9
PixArt-Sigma	dot-product	dot-product	43.2
	l2-norm	dot-product	43.2
	cosine	dot-product	41.2
	dot-product	mse	**44.1**
	dot-product	iou-like	43.4