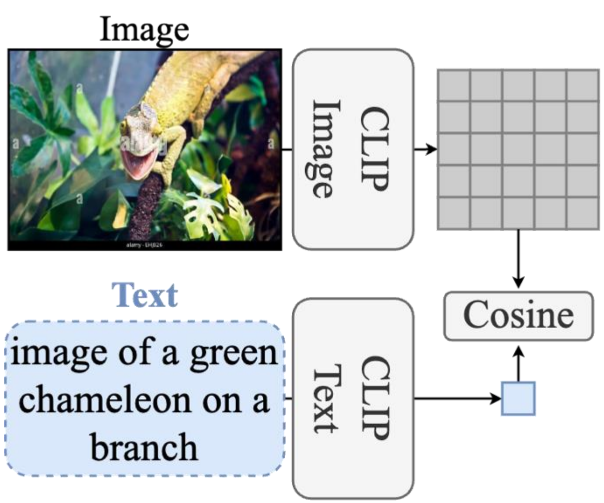
TTD: Text-Tag Self-Distillation Enhancing Image-Text Alignment in CLIP to Alleviate Single Tag Bias

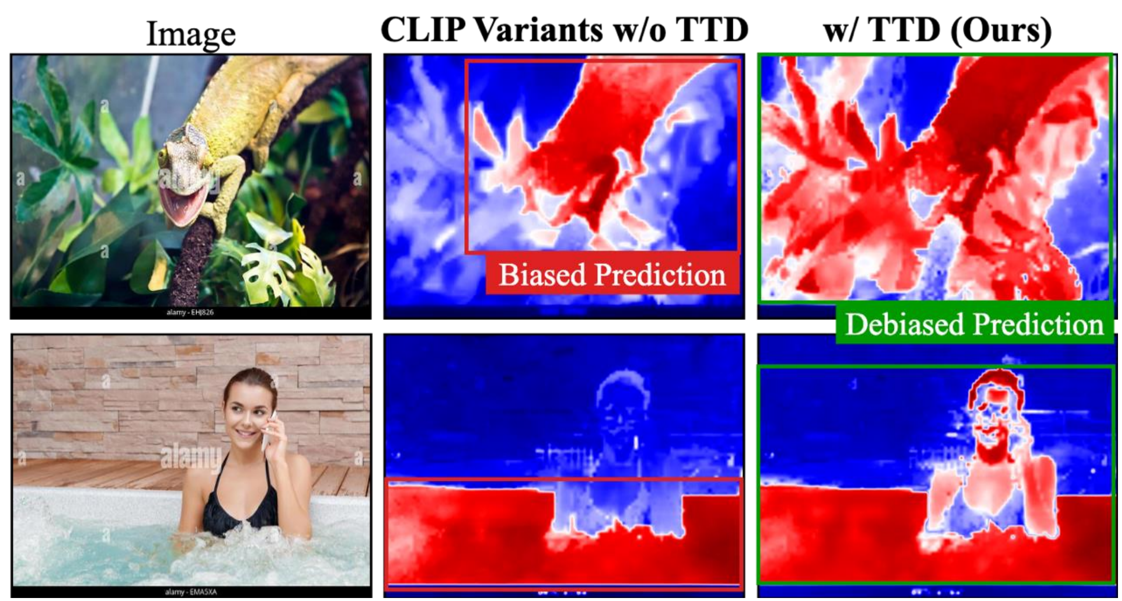
Sanghyun Jo, Soohyun Ryu*, Sungyub Kim, Eunho Yang, Kyungsu Kim*

Problem: Single Tag Bias in CLIP-based Models

- **Single tag bias** manifests as a disproportionate focus on a singular tag (word) while neglecting other pertinent tags.

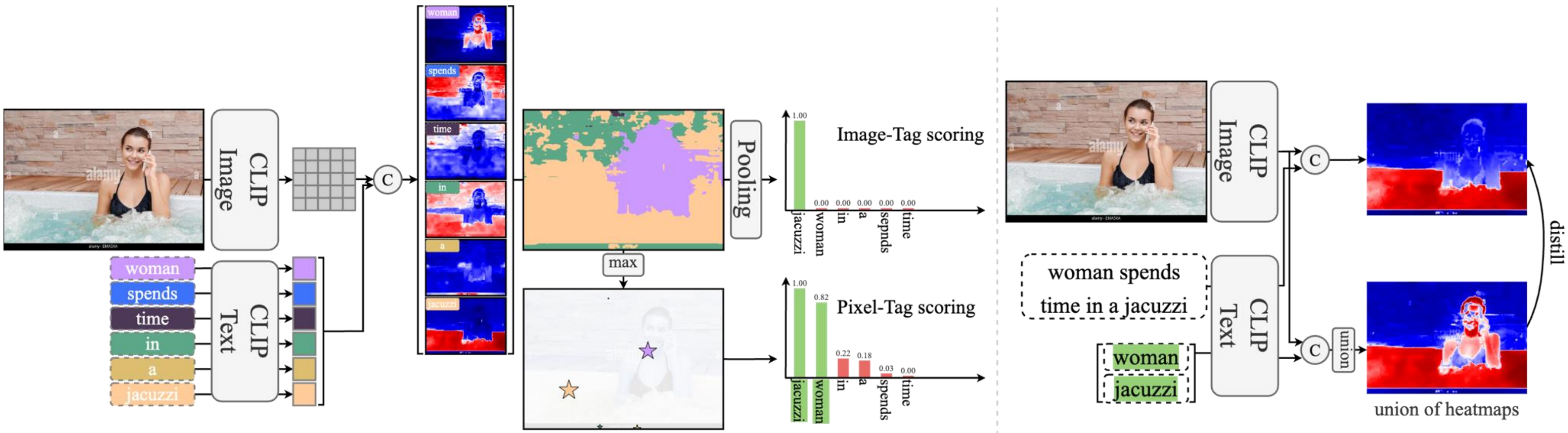
a green chameleon on a branch





Method: Text-Tag Self-Distillation (TTD)

- Two-step fine-tuning approach that effectively mitigates single tag bias
  - Enables models to recognize all relevant tags
- Model-agnostic
- No external supervision is required

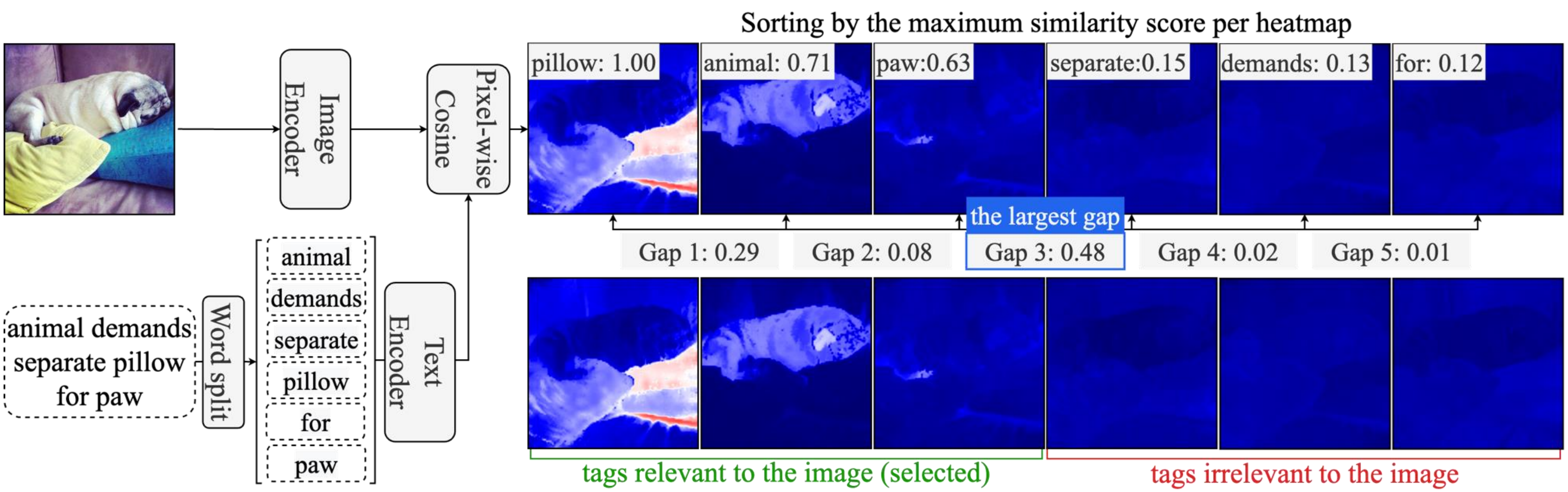


(a) Tag selection by pixel-tag scoring

(b) Text-tag self-distillation

Method: Text-Tag Self-Distillation (TTD)

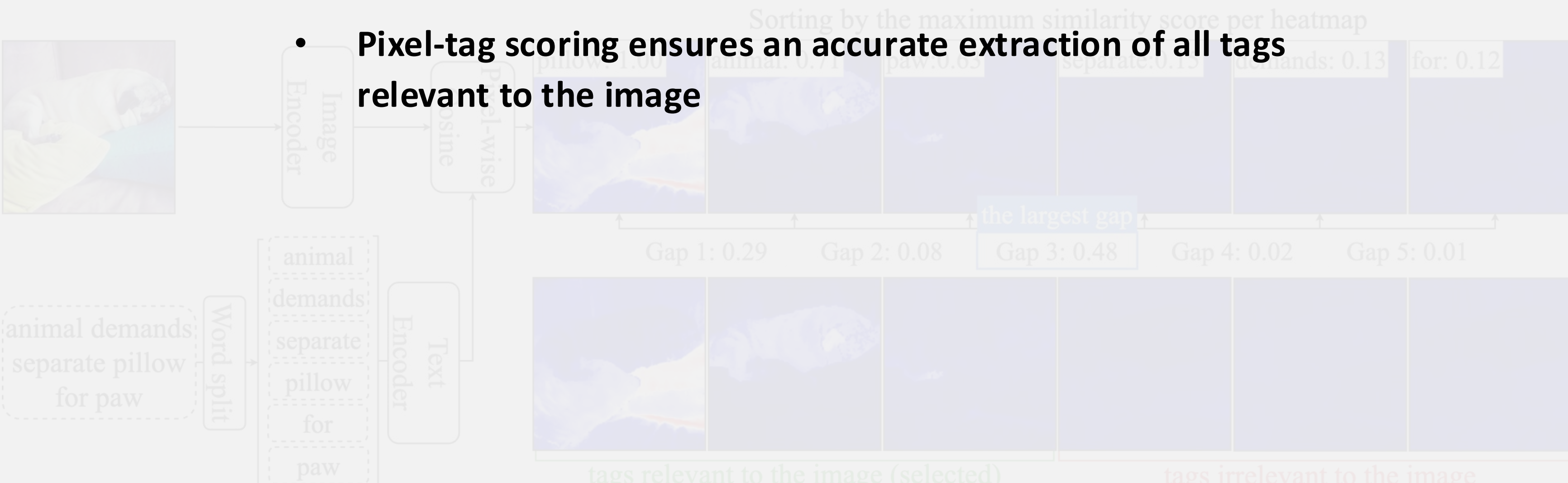
- **Step1) Tag Selection by Pixel-Tag Scoring**
  - **Identify which tags in the text are relevant to the image**
  - Move from global to pixel-level embedding
  - Score tags based on their correlation with specific image regions



Method: Text-Tag Self-Distillation (TTD)

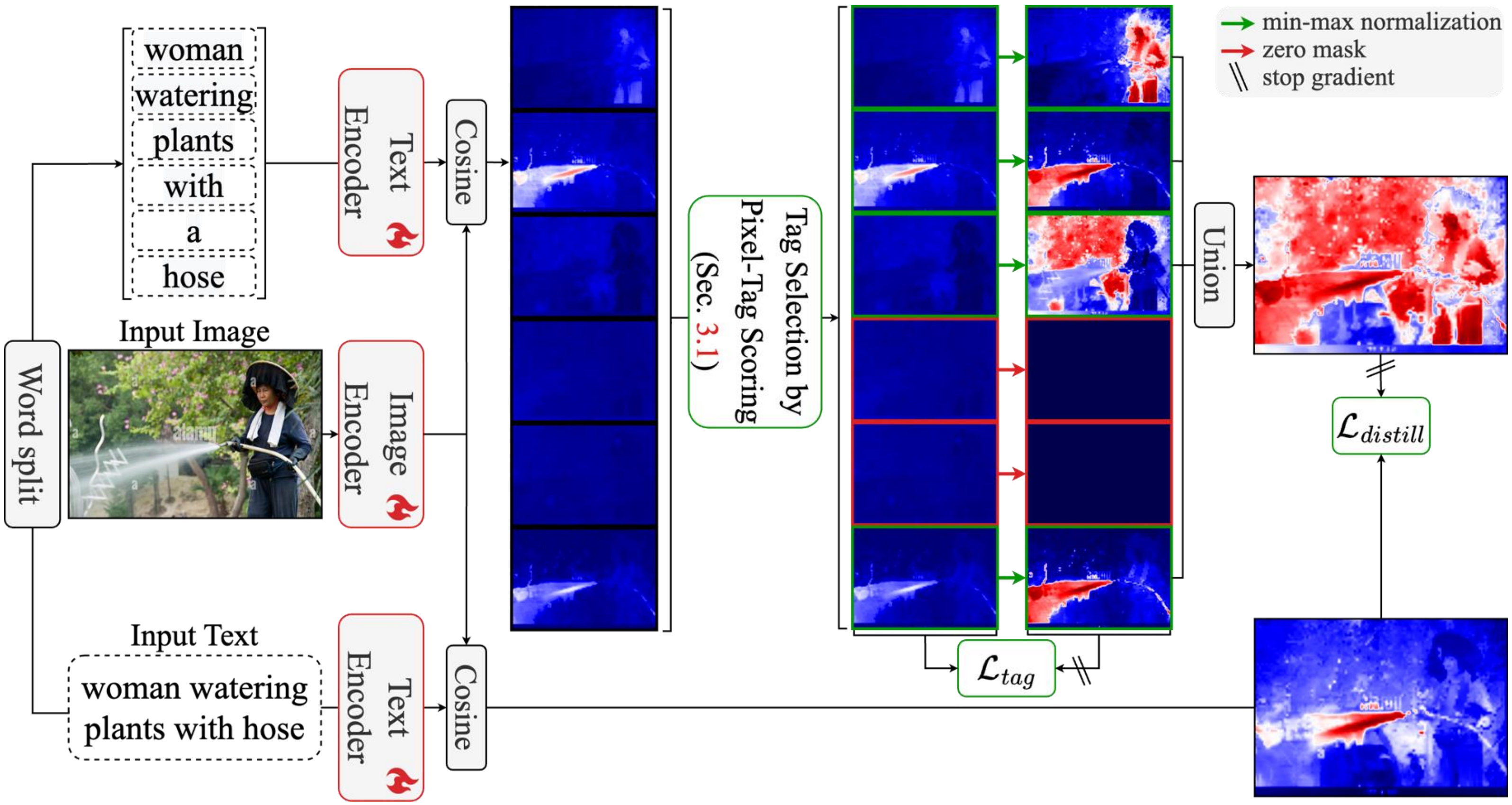
- Step1) Tag Selection by Pixel-Tag Scoring
  - Identify which tags in the text are relevant to the image
  - Move from global to pixel-level embedding
  - Score tags based on their correlation with specific image regions

- **Why do we do this?**
  - **Global embedding often overemphasizes the dominant tag**
  - **Pixel-tag scoring ensures an accurate extraction of all tags relevant to the image**



Method: Text-Tag Self-Distillation (TTD)

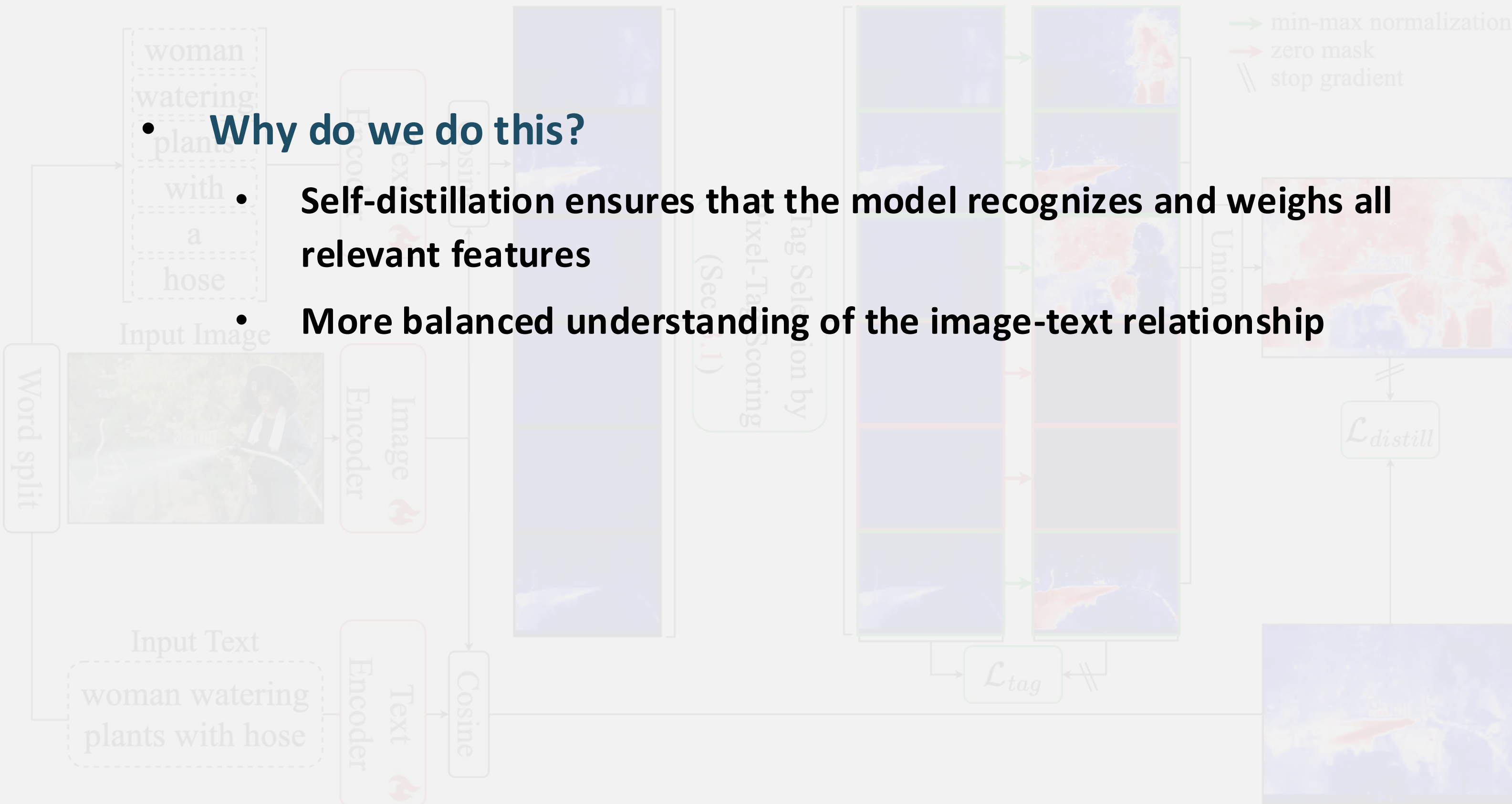
- **Step2) Text-Tag Self-Distillation**
  - Generate a composite mask using multiple tags
  - Self-distillation ensures the model learns to align the text with the composite mask



Method: Text-Tag Self-Distillation (TTD)

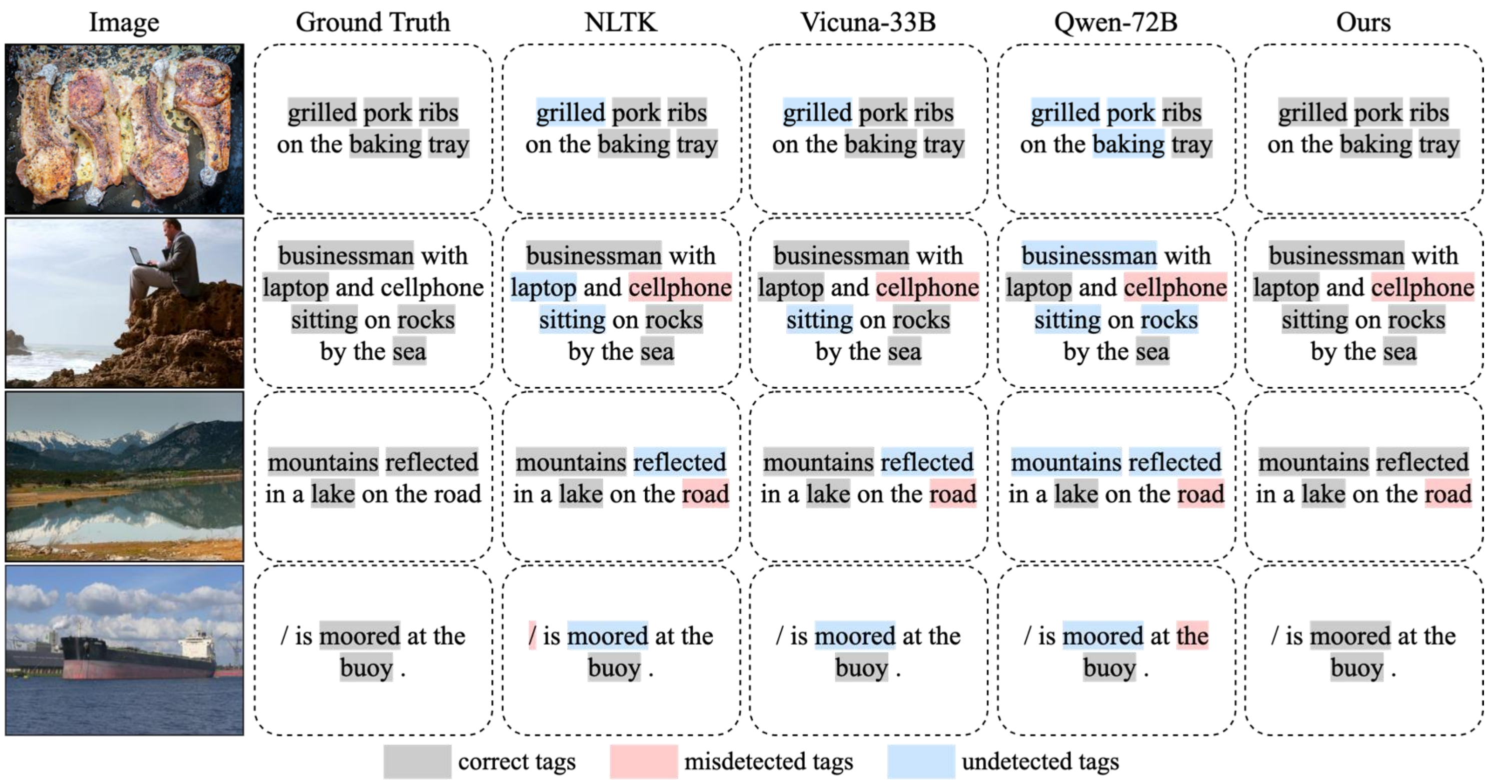
- Step2) Text-Tag Self-Distillation
  - Generate a composite mask using multiple tags
  - Self-distillation ensures balanced image-text alignment

- **Why do we do this?**
  - **Self-distillation ensures that the model recognizes and weighs all relevant features**
  - **More balanced understanding of the image-text relationship**



Experiments: Multi-Tag Selection

- **Tag selection using external models**
  - Extracting Image-irrelevant Tags (red)
  - Overlooking Image-relevant Tags (blue)



Experiments: Multi-Tag Selection

- **Quantitative results**
  - s_{image} tends to focus on a single dominant tag
  - F1 Score: 82.8%
  - Significant improvement over external models or baseline scoring methods

Table 2: Multi-Tag Selection. The best results are **bold** and the second best results are underlined. P, precision; R, recall; F1, F1 score.

(a) Comparison based on the use of NLP models.

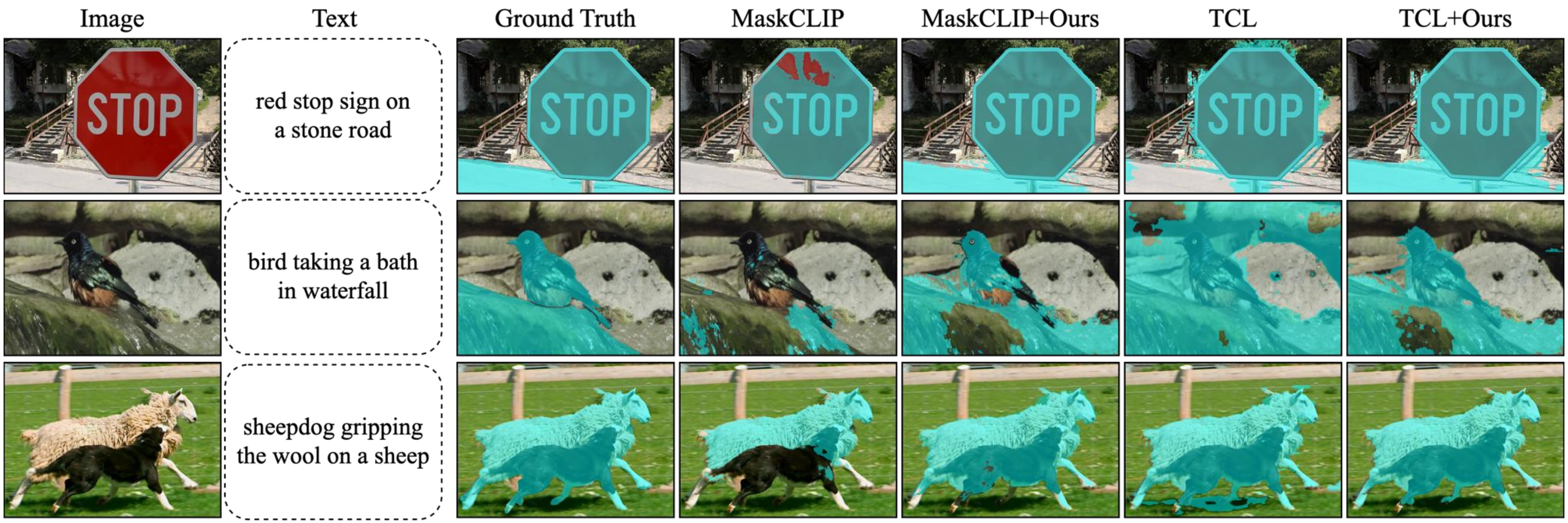
Method	P	R	F1	Acc
NLTK	59.8	**83.7**	69.8	79.6
Vicuna-7B	44.1	71.0	54.4	70.9
Vicuna-33B	52.7	70.7	60.4	75.9
Qwen-72B	69.3	56.2	62.1	80.9
$s_{\text{pixel}}^{\text{ours}}$ (Eq. (2))	82.9	74.5	78.5	88.6
+ TTD (Ours)	**88.3**	78.0	**82.8**	**91.0**

(b) Comparison based on scoring methods.

Scoring	P	R	F1	Acc	mAP
s_{image} (Eq. (1))	**92.5**	28.6	43.7	79.5	83.2
s_{text} (Eq. (7))	85.6	29.7	44.1	79.0	82.1
$s_{\text{image}} + s_{\text{text}}$	85.5	45.1	59.0	82.6	84.5
$s_{\text{pixel}}^{\text{ours}}$ (Eq. (2))	82.9	74.5	78.5	88.6	90.3
+ TTD (Ours)	88.3	**78.0**	**82.8**	**91.0**	**93.7**

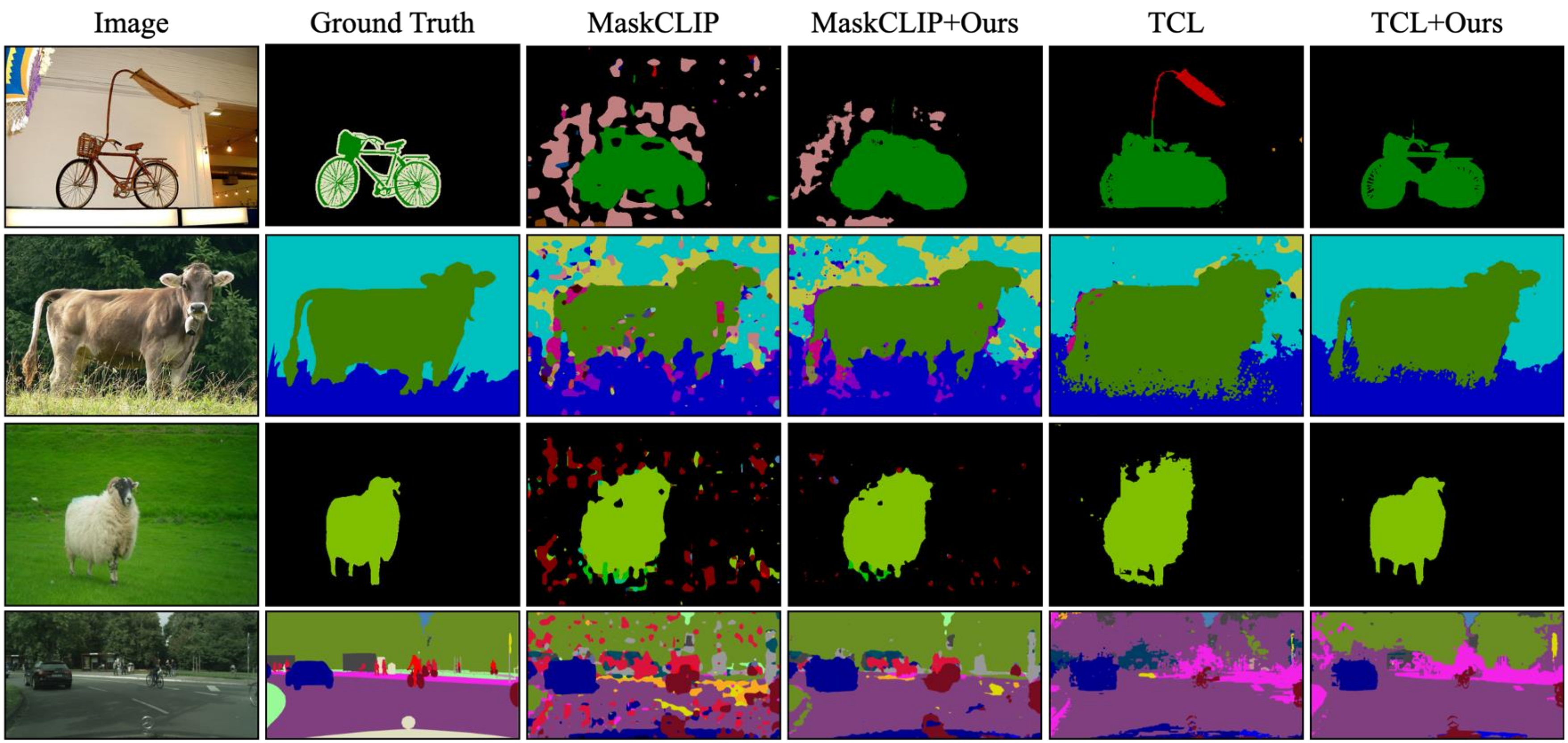
Experiments: Text-Level Segmentation

- **CaptionIoU improvement: +9.2% for MaskCLIP, +5.1% for TCL**



Method	CaptionIoU(%)	mFPR	mFNR
MaskCLIP [44]	41.0	0.179	0.411
+ TTD (Ours)	50.2 (+9.2)	0.256	0.242
TCL [4]	60.4	0.199	0.198
+ TTD (Ours)	65.5 (+5.1)	0.163	0.182

Experiments: Open-Vocabulary (Tag-Level) Segmentation



Experiments: Open-Vocabulary (Tag-Level) Segmentation

- **mIoU improvement: +8.5% for MaskCLIP, +3.5% for TCL**

Table 4: Open-Vocabulary Semantic Segmentation. The methods are all trained only with image and text data, without additional annotations or external models. We use ViT-B/16 as the backbone for all methods. $\mathcal{L}$, external language models.

Method	Train datasets	$\mathcal{L}$	VOC	Context	Object	Stuff	City	ADE	Avg.
GroupViT [38]	CC12M+YFCC	✓	51.1	19.0	27.9	15.4	11.6	9.4	22.4
ViewCo [33]	CC12M+YFCC	✓	52.4	23.0	23.5	-	-	-	-
CoCu [37]	CC3M+CC12M+COCO	✓	51.4	23.6	22.7	15.2	22.1	12.3	24.6
OVSegmentor [39]	CC4M [39]	✓	53.8	20.4	25.1	-	-	-	-
TagAlign [24]	CC12M	✓	53.9	33.5	33.3	**25.3**	<u>27.5</u>	**17.3**	<u>31.8</u>
ReCo [35]	ImageNet1K	✗	25.1	19.9	15.7	14.8	21.1	11.2	18.0
ZeroSeg [6]	ImageNet1K	✗	40.8	20.4	20.2	-	-	-	-
ViL-Seg [25]	CC12M	✗	37.3	18.9	18.1	-	-	-	-
SimSeg [41]	CC3M+CC12M	✗	<u>57.4</u>	26.2	29.7	-	-	-	-
SegCLIP [27]	CC12M+COCO	✗	52.6	24.7	26.5	16.1	11.2	8.8	23.3
MaskCLIP [44]	-	✗	29.3	21.1	15.5	14.7	21.6	10.4	19.0
+ TTD (Ours)	CC3M+CC12M	✗	43.1 (+13.8)	31.0 (+9.9)	26.5 (+11.0)	19.4 (+4.7)	**32.0** (+10.4)	12.7 (+2.3)	27.5 (+8.5)
TCL [4]	CC3M+CC12M	✗	55.0	<u>33.8</u>	31.6	22.4	24.0	15.6	30.4
+ TTD (Ours)	CC3M+CC12M	✗	**61.1** (+6.1)	**37.4** (+3.6)	**37.4** (+5.8)	<u>23.7</u> (+1.3)	27.0 (+3.0)	<u>17.0</u> (+1.4)	**33.9** (+3.5)

Ablation

- Using both distillation and auxiliary losses yields best performance (+6.1% mIoU)
- Pixel-tag scoring outperforms standard tag selection methods (higher F1 and mIoU)

(a) Effect of Loss Terms.

$\mathcal{L}_{distill}$ (Eq. (4))	$\mathcal{L}_{tag}(Eq.\ (5))$	CaptionIoU	mIoU
✗	✗	60.4	55.0
✗	✓	60.7 (+0.3)	58.5 (+3.5)
✓	✗	63.6 (+3.2)	60.8 (+5.8)
✓	✓	**65.5** (+5.1)	**61.1** (+6.1)

(b) Effect of Tagging Method.

Method	CaptionIoU	mIoU
Baseline [4]	60.4	55.0
NLTK [26]	61.8	56.5
s_{image} (Eq. (1))	56.3	52.5
$s_{\text{pixel}}^{\text{ours}}$ (Eq. (2))	**65.5**	**61.1**

Conclusion

- **Text-Tag Self-Distillation (TTD) addresses single tag bias in CLIP-based models**
- **Model-agnostic with no external data or model required**
- **Enhanced performance across three tasks: multi-tag selection, text-level segmentation, and open-vocabulary segmentation**

Thank you !