Multi-Label Test-Time Adaptation with Bound Entropy Minimization

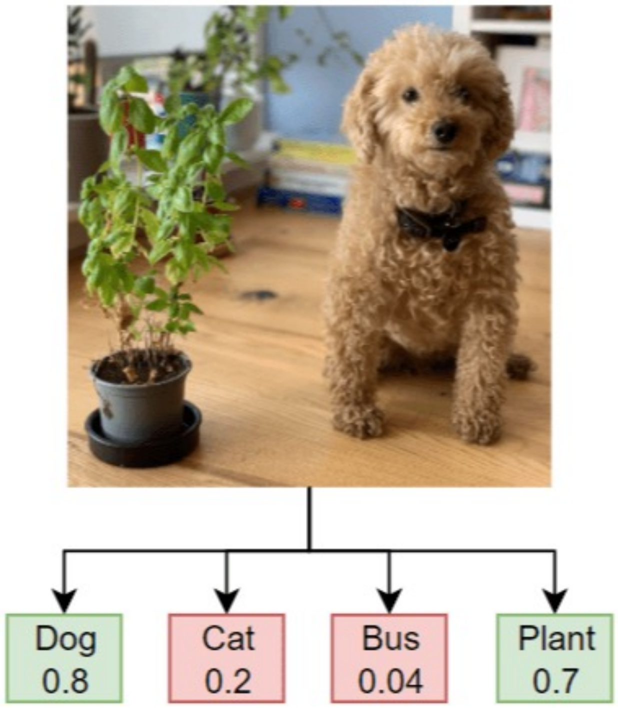
Xiangyu Wu[1,2], Feng Yu[1], Qing-Guo Chen[2] , Yang Yang[1*], Jianfeng Lu[1*]

[1]NJUST [2]Alibaba

ICLR2025

CONTENT

Background: Multi-Label Recognition



Label Set: {Dog, Cat, Bus, Plant}

Task type	Train input	Train label	Test input
Full-shot	All categories All images	All labeled	Same categories
Few-shot	All categories Few images	All labeled	Same categories
Zero-shot	Normal or No categories No images	All labeled	Novel categories
ML-TTA	No Access Training data and Source Model	No Access	Same categories

计算机科学与软件学院 人工智能学院 工程学院
SCHOOL OF SOFTWARE SCHOOL OF ARTIFICIAL INTELLIGENCE SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Background:

- The TTA (Test Time Augmentation) technique is mainly aimed at multi-class classification tasks, and it increases the probability of the most confident label by minimizing the entropy.
- In the multi-label scenario, since the number of labels for each image varies, focusing only on the label with the highest probability will lead to a decrease in the adaptability of other positive labels.

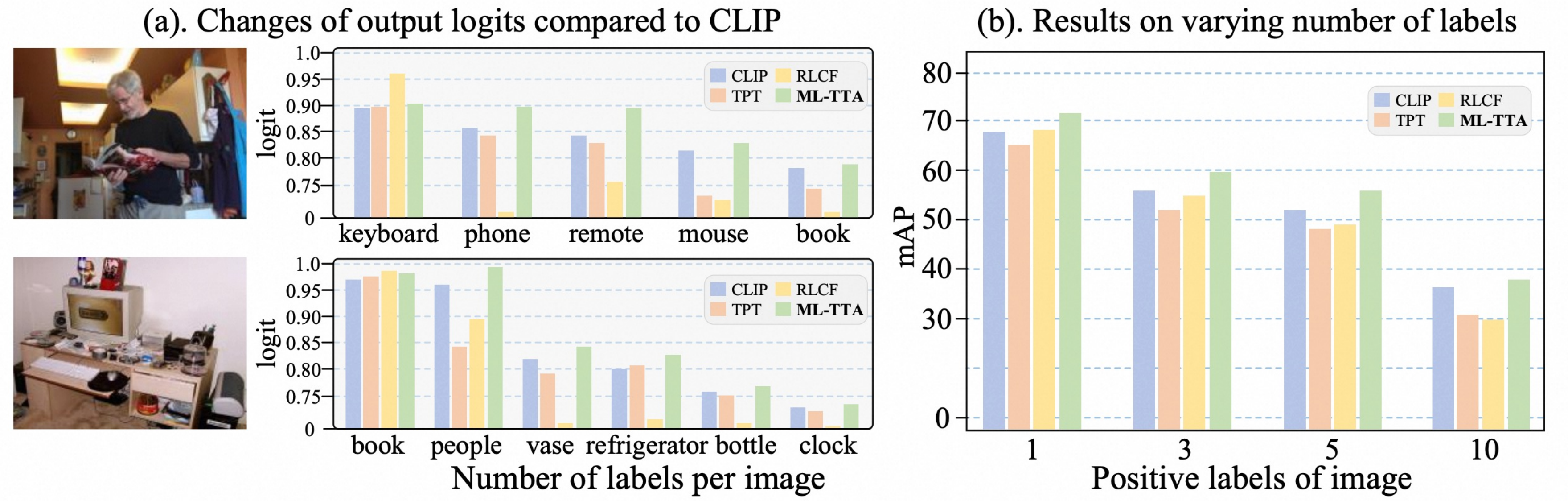


Figure 1: (a). Compared to CLIP (Radford et al., 2021), ML–TTA increases all positive label logits simultaneously, while others focus only on *top-1* class. (b). Comparison of various methods on images with varying numbers. Compared to CLIP, as the number of labels per image rises, the adaptability of TPT (Shu et al., 2022) and RLCF (Zhao et al., 2024a) in handling multi-label images shows a marked decrease.

计算机科学与软件学院 SCHOOL OF SOFTWARE
人工智能学院 SCHOOL OF ARTIFICIAL INTELLIGENCE 工程学院
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Bound Entropy Minimization (BEM)

- Pair each augmented view of the image with a text caption.
- Extract the textual label from the caption as the *"strong label set"* for the description and the *"weak label set"* for the augmented view.
- Bind the *strong label set* and the *weak label set* into single labels respectively, and optimize the view prompt and the caption prompt to improve the confidence of the top-k labels.

Multi-Label Test-Time Adaptation (ML-TTA) framework

- Taking the TPT method as the starting point, combine it with the BEM objective for adaptation during multi-label testing.
- Reduce noise by filtering out views and captions with high entropy (low confidence).
- Experiments on different model architectures, prompt initializations, and label scenarios have demonstrated the effectiveness of the ML-TTA framework.

Bound Entropy Minimization (BEM)

Proposition 1：

- Consider the output logits of a confident view. Assume that $s_1 > s_2 > \ldots > s_L$
- The entropy loss H decreases as s_1 increases, and increases as the sum s_{test} of the remaining logits decreases.
- That is, the entropy loss tends to increase the probability of the most confident label while reducing the sum of the probabilities of the other labels.

$$\nabla_{s_1} H = \frac{\partial H}{\partial s_1} < 0 \quad and \quad \nabla_{s_{rest}} H = \frac{\partial H}{\partial S_{rest}} > 0.$$

- Consider the gradient descent update for one step, and it can be deduced that:

$$s_1^{(t+1)} = s_1^{(t)} - \alpha \nabla_{s_1} H \text{ and } S_{rest}^{(t+1)} = S_{rest}^{(t)} - \alpha \nabla_{S_{rest}} H$$

Proposition 1 explains why the traditional entropy minimization method is not applicable to the multi-label scenario: It will only increase the probability of the most confident label while ignoring other positive labels.

Bound Entropy Minimization (BEM)

Proposition 2：

- Consider the output logits of a confident view. Assume that $s_1 > s_2 > \dots > s_L$
- Define the modified logits as ***s'***, where $s'_i = a_i + s_i \ (i <= k)$ and $s'_i = s_i \ (i > k)$, and a_i is a constant.
- For the modified logits ***s'***, define the modified probability ***p'=Softmax(s')*** and the modified entropy $H' = -\Sigma\, p'_i \log p'_i$.
- The gradient properties of the modified entropy ***H'*** are as follows:

$$\nabla_{s_i} H' = \frac{\partial H'}{\partial s_i} < 0, \quad \forall i \leq k \quad and \quad \nabla_{s_{rest}} H' = \frac{\partial H'}{\partial S_{rest}} > 0.$$

- Similarly, after one-step gradient descent optimization, the predicted probabilities of all the ***top-k*** predicted labels will increase further.

Proposition 2 leads to the BEM objective:

- By regarding the top-k predicted labels as a single entity, it effectively addresses the limitations of entropy minimization in the multi-label scenario.
- The BEM objective encourages the model to increase the probabilities of multiple top-k labels simultaneously, thus enabling it to better adapt to multi-label data.

Method: Multi-Label Test-Time Adaptation (ML-TTA)

Multi-Label Test-Time Adaptation (ML-TTA)

1. View-caption Construction:

- Define ***view prompt*** and ***caption prompt***: *a photo of a [CLS]*
- Input an image x^{test}, and obtain N augmented views, each view is assigned with a retrieved paired caption:

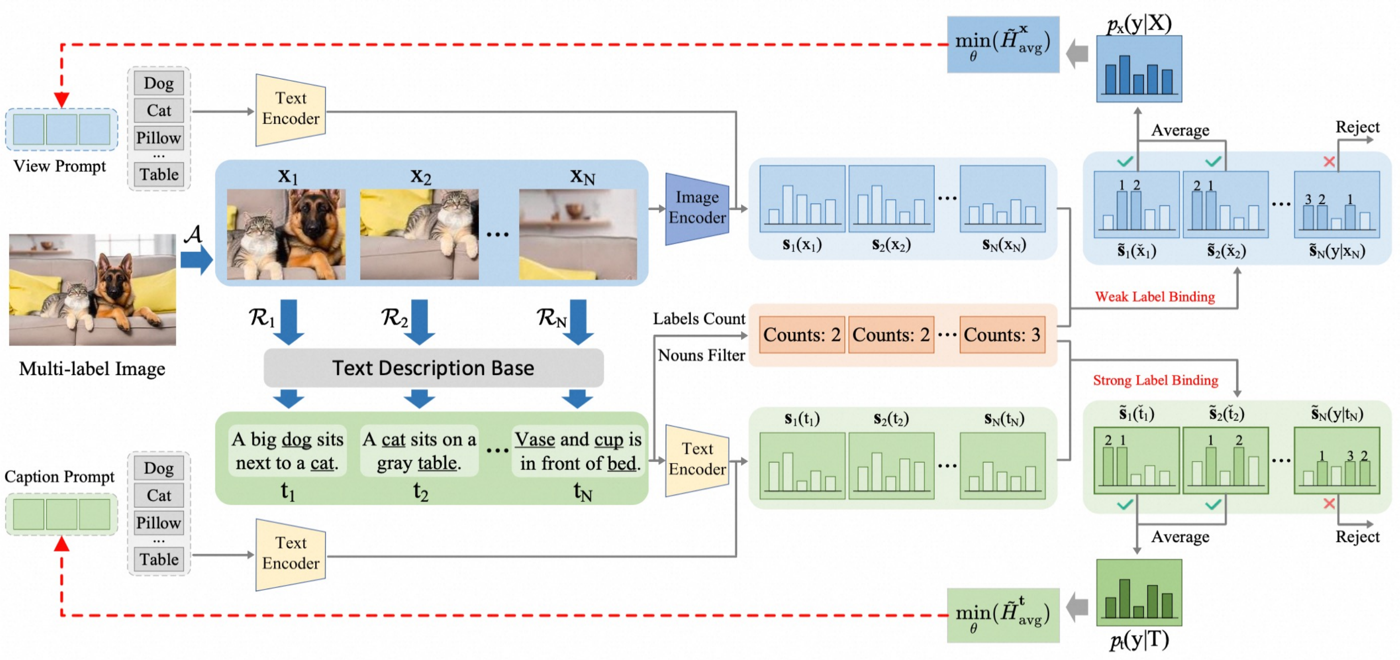
$$X^{\text{test}} = \{\mathbf{x}_i^{\text{test}} \mid \mathbf{x}_i^{\text{test}} = \mathcal{A}_i(\mathbf{x}^{\text{test}})\}_{i=1}^{N}$$

$$T^{\text{test}} = \{\mathbf{t}_i^{\text{test}} \mid \mathbf{t}_i^{\text{test}} = \mathcal{R}_i(\mathbf{x}_i^{\text{test}})\}_{i=1}^{N}$$

- Compute logits:

$$s_{ij}^{\mathbf{x}^{\text{test}}} = \langle \text{Enc}^{\text{I}}(\mathbf{x}_i^{\text{test}}), \text{Enc}^{\text{T}}(\mathbf{v}_j) \rangle$$

$$s_{ij}^{\mathbf{t}^{\text{test}}} = \langle \text{Enc}^{\text{T}}(\mathbf{t}_i^{\text{test}}), \text{Enc}^{\text{T}}(\mathbf{c}_j) \rangle$$



Method: Multi-Label Test-Time Adaptation (ML-TTA)

Multi-Label Test-Time Adaptation (ML-TTA)

2. Label Binding

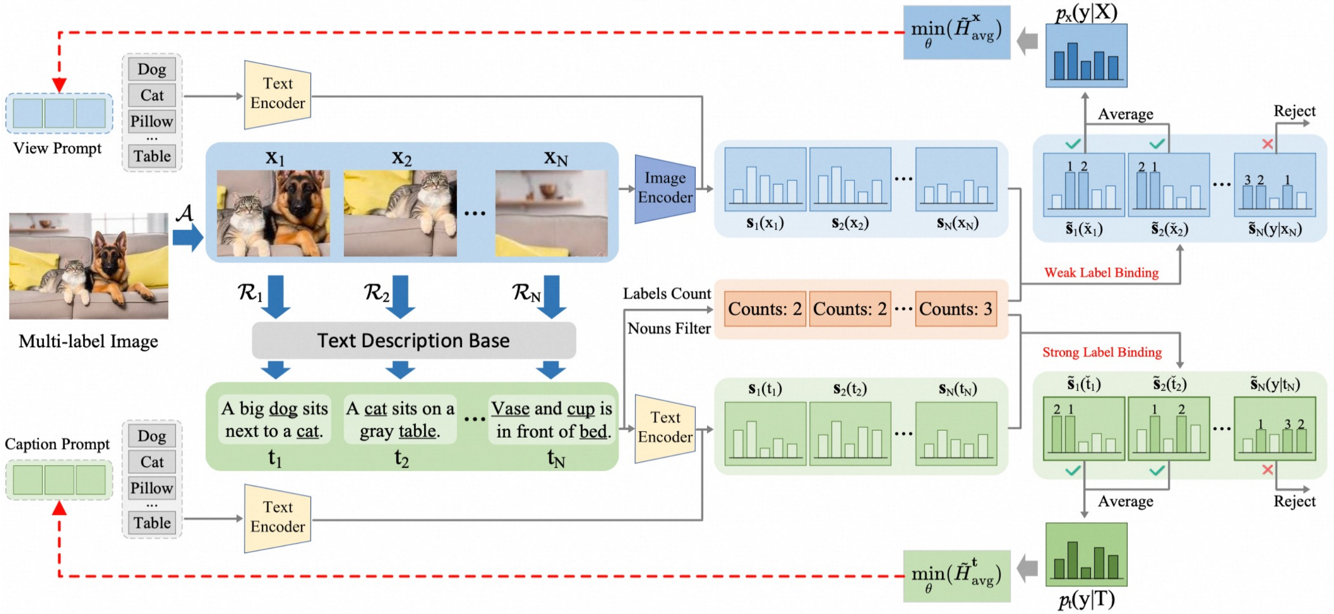
- Extract textual labels from caption as strong label set of that caption and as weak label set of that view, with same size k
- Bind the top-k labels of each view into a single label to make their logits equal, and do the same for the captions.

$$\tilde{s}_{ij}^{\mathbf{x}^{\text{test}}} = ((m_i^{\mathbf{x}^{\text{test}}} - s_{ij}^{\mathbf{x}^{\text{test}}}) + s_{ij}^{\mathbf{x}^{\text{test}}}) \cdot \mathbb{I}(\text{Rank}_{(s_{ij}^{\mathbf{x}^{\text{test}}}, \mathbf{s}_i^{\mathbf{x}^{\text{test}}})} \le k^{\mathbf{x}_i^{\text{test}}}) + s_{ij}^{\mathbf{x}^{\text{test}}} \cdot \mathbb{I}(\text{Rank}_{(s_{ij}^{\mathbf{x}^{\text{test}}}, \mathbf{s}_i^{\mathbf{x}^{\text{test}}})} > k^{\mathbf{x}_i^{\text{test}}}),$$

$$\tilde{s}_{ij}^{\mathbf{t}^{\text{test}}} = ((m_i^{\mathbf{t}^{\text{test}}} - s_{ij}^{\mathbf{t}^{\text{test}}}) + s_{ij}^{\mathbf{t}^{\text{test}}}) \cdot \mathbb{I}(\text{Rank}_{(s_{ij}^{\mathbf{t}^{\text{test}}}, \mathbf{s}_i^{\mathbf{t}^{\text{test}}})} \le k^{\mathbf{t}_i^{\text{test}}}) + s_{ij}^{\mathbf{t}^{\text{test}}} \cdot \mathbb{I}(\text{Rank}_{(s_{ij}^{\mathbf{t}^{\text{test}}}, \mathbf{s}_i^{\mathbf{t}^{\text{test}}})} > k^{\mathbf{t}_i^{\text{test}}}),$$

- Filter out the views and captions with low entropy (high confidence) through confidence threshold, and calculate the average entropy respectively:

$$\tilde{H}_{\text{avg}}^{\check{\mathbf{x}}^{\text{test}}} = \frac{1}{\tau N}\sum_{i=1}^{\tau N}\left(-\sum_{l=1}^{L} p(y=l|\check{\mathbf{x}}_i^{\text{test}})\log(p(y=l|\check{\mathbf{x}}_i^{\text{test}}))\right)$$



- Total Objective：

$$\tilde{H}_{\text{BEM}} = \tilde{H}_{\text{avg}}^{\check{\mathbf{x}}^{\text{test}}} + \tilde{H}_{\text{avg}}^{\check{\mathbf{t}}^{\text{test}}}.$$

Multi-Label Test-Time Adaptation (ML-TTA)

Algorithm 1: Label Binding Algorithm

Input: Logits $\mathbf{s}_i$ before label binding and the size of weak label set $k^{\mathbf{x}_i}$.
Output: Modified logits $\tilde{\mathbf{s}}_i$ after label binding.

$m_i = \max_j s_{ij}$;
for $j = 1$ **to** L **do**
  $a_{ij} = \text{detach}\,(m_i - s_{ij})$ ▷ Detach from gradient. ;
  if $\text{Rank}_{(s_{ij}, \mathbf{s}_i)} \leq k^{\mathbf{x}_i}$ **then**
    $\tilde{s}_{ij} = a_{ij} + s_{ij}$ ▷ Bind s_{ij} if j-th label is in highest *top*-$k^{\mathbf{x}_i}$ predicted labels. ;
  end if
  else
    $\tilde{s}_{ij} = s_{ij}$;
  end if
end for
$\tilde{\mathbf{s}}_i = (\tilde{s}_{i0}, \tilde{s}_{i1}, \cdots, \tilde{s}_{iL})$

Experiments

Results on different architectures.

	Method	*Epsdoic*	COCO2014	COCO2017	VOC2007	VOC2012	NUSWIDE	Average
RN-50	CLIP [ICML 2022]	✓	47.53	47.32	75.91	74.25	41.53	57.31
	DMN [CVPR 2024]	×	44.54	44.18	74.87	74.13	41.32	55.81
	TDA [CVPR 2024]	×	48.91	49.11	76.64	75.12	42.34	58.42
	TPT [NeurIPS 2022]	✓	48.52	48.51	75.54	73.92	41.97	57.69
	DIffTPT [ICCV 2023]	✓	48.56	48.67	75.89	74.13	41.33	57.72
	RLCF [ICLR 2024]	✓	36.87	36.73	65.75	64.73	29.83	46.78
	ML–TTA (Ours)	✓	**51.58**	**51.39**	**78.62**	**76.63**	**42.53**	**60.15**
RN-101	CLIP [ICML 2022]	✓	48.83	48.15	76.72	74.21	41.93	57.97
	DMN [CVPR 2024]	×	46.28	45.44	76.82	75.32	42.71	57.31
	TDA [CVPR 2024]	×	50.19	49.78	78.12	77.13	43.13	59.67
	TPT [NeurIPS 2022]	✓	49.71	48.89	74.82	73.39	43.10	57.98
	DIffTPT [ICCV 2023]	✓	49.45	49.19	74.98	74.31	42.93	58.17
	RLCF [ICLR 2024]	✓	40.53	39.79	71.21	69.63	31.77	50.59
	ML–TTA (Ours)	✓	**52.92**	**52.24**	**78.72**	**78.13**	**43.62**	**61.13**
ViT-B/32	CLIP [ICML 2022]	✓	50.31	50.15	77.18	76.85	42.90	59.48
	DMN [CVPR 2024]	×	49.32	48.13	77.42	76.60	43.41	58.98
	TDA [CVPR 2024]	×	51.23	51.49	77.62	77.12	44.13	60.32
	TPT [NeurIPS 2022]	✓	48.12	48.63	74.21	71.93	43.63	57.30
	DIffTPT [ICCV 2023]	✓	48.73	49.19	74.50	72.98	43.42	57.76
	RLCF [ICLR 2024]	✓	50.28	49.59	77.12	76.83	43.29	59.42
	ML–TTA (Ours)	✓	**52.83**	**52.99**	**78.70**	**77.97**	**44.12**	**61.32**
ViT-B/16	CLIP [ICML 2022]	✓	54.42	54.13	79.58	79.25	45.65	62.61
	DMN [CVPR 2024]	×	52.52	52.37	79.83	79.67	46.27	62.13
	DART [AAAI 2024]	×	54.73	54.68	79.91	78.56	45.91	62.76
	TDA [CVPR 2024]	×	55.21	55.46	80.12	79.92	46.72	63.49
	TPT [NeurIPS 2022]	✓	53.32	54.20	77.54	77.39	46.15	61.72
	DIffTPT [ICCV 2023]	✓	53.91	54.15	77.93	77.24	46.13	61.87
	RLCF [ICLR 2024]	✓	54.21	54.43	79.29	79.26	43.18	62.07
	ML–TTA (Ours)	✓	**57.52**	**57.49**	**81.28**	**81.13**	**46.55**	**64.80**

Results on different prompt initialization.

Table 2: Comparison with SOTAs on adapting multi-label instances with different prompt initialization.

	Methods	*Epsdoic*	COCO2014	COCO2017	VOC2007	VOC2012	NUSWIDE	Average
CoOp	CoOp [IJCV2022]	✓	56.12	56.35	79.14	77.85	46.74	63.24
	TDA [CVPR 2024]	×	56.93	57.15	80.20	78.58	47.82	64.13
	TPT [NeurIPS 2022]	✓	55.35	55.23	79.72	77.85	47.27	63.08
	DIffTPT [ICCV 2023]	✓	55.30	55.47	79.86	77.61	47.13	63.07
	RLCF [ICLR 2024]	✓	56.72	56.18	80.15	78.24	47.62	63.78
	ML–TTA (Ours)	✓	**59.68**	**59.33**	**83.17**	**81.36**	**48.12**	**66.33**
Maple	Maple [CVPR2023]	✓	62.18	62.35	85.34	84.79	48.42	68.62
	TDA [CVPR 2024]	×	63.25	63.64	85.76	84.15	49.55	69.27
	TPT [NeurIPS 2022]	✓	63.36	63.75	85.04	83.92	48.90	69.01
	DIffTPT [ICCV 2023]	✓	62.93	63.14	85.15	83.78	48.81	68.76
	RLCF [ICLR 2024]	✓	62.84	62.90	85.35	85.28	49.37	69.15
	ML–TTA (Ours)	✓	**64.75**	**64.86**	**86.40**	**85.69**	**50.21**	**70.38**

Results on different label counts.

Table 3: Results on different label counts.

Methods	{1,2}	{3,4}	{5,6,7}	{>=8}
CLIP [ICML 2022]	62.76	<u>55.41</u>	<u>49.89</u>	<u>41.07</u>
TPT [NeurIPS 2022]	62.88	53.05	45.57	37.43
DiffTPT [ICCV 2023]	61.97	52.67	44.32	36.89
RLCF [ICLR 2024]	<u>66.01</u>	51.65	43.32	35.08
ML–TTA (Ours)	**67.14**	**57.59**	**51.68**	**41.32**

計算機科學與軟件學院
人工智能學院 工程學院
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Further analysis

Table 6: Comparison with binary cross-entropy loss.

Methods	RN50		ViT-B/16	
	COCO2014	VOC2007	COCO2014	VOC2007
CLIP	47.53	75.91	54.42	79.58
VP+CP+BCE	48.39	75.75	54.51	78.59
VP+CP+BEM	**51.58**	**78.62**	**57.52**	**81.28**

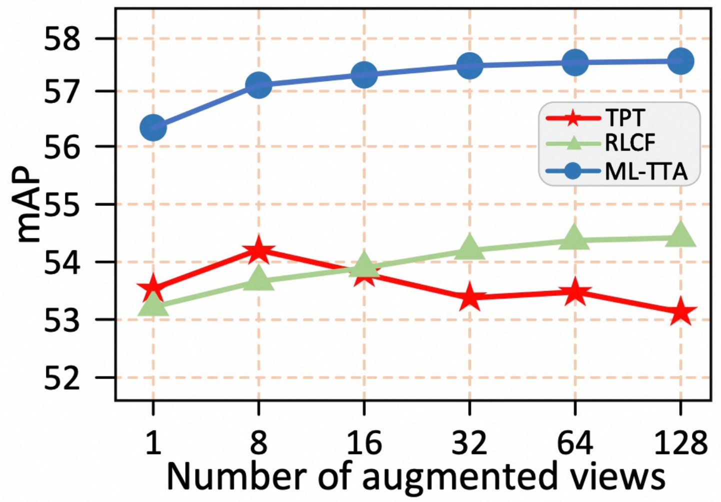


Figure 3: Results on different number of views.

Table 7: Results on different numbers of retrieved captions.

Datasets		CLIP	TPT	1	2	4	8	16	32	64	128
RN50	COCO2014	47.53	48.52	51.35	51.37	51.41	51.49	51.58	**51.59**	51.55	51.48
	VOC2007	75.91	75.54	78.29	78.33	78.48	78.54	**78.61**	78.59	78.53	78.42
ViT-B/16	COCO2014	54.42	53.32	57.23	57.33	57.41	57.48	57.49	57.52	57.55	**57.58**
	VOC2007	79.58	77.54	81.06	81.12	81.21	81.24	**81.28**	81.19	81.15	80.98

THANK YOU！