NEURAL INFORMATION PROCESSING SYSTEMS

Text-Guided Attention is All You Need for Zero-Shot Robustness in Vision-Language Models

Lu Yu Haiyang Zhang Changsheng Xu

TIANJIN UNIVERSITY OF TECHNOLOGY 天津理工大学 1979

CASIA

Adversarial Attack

Deep neural networks have been found to be vulnerable to adversarial examples.

Original Example
“golden retriever”

+

Adversarial Perturbation

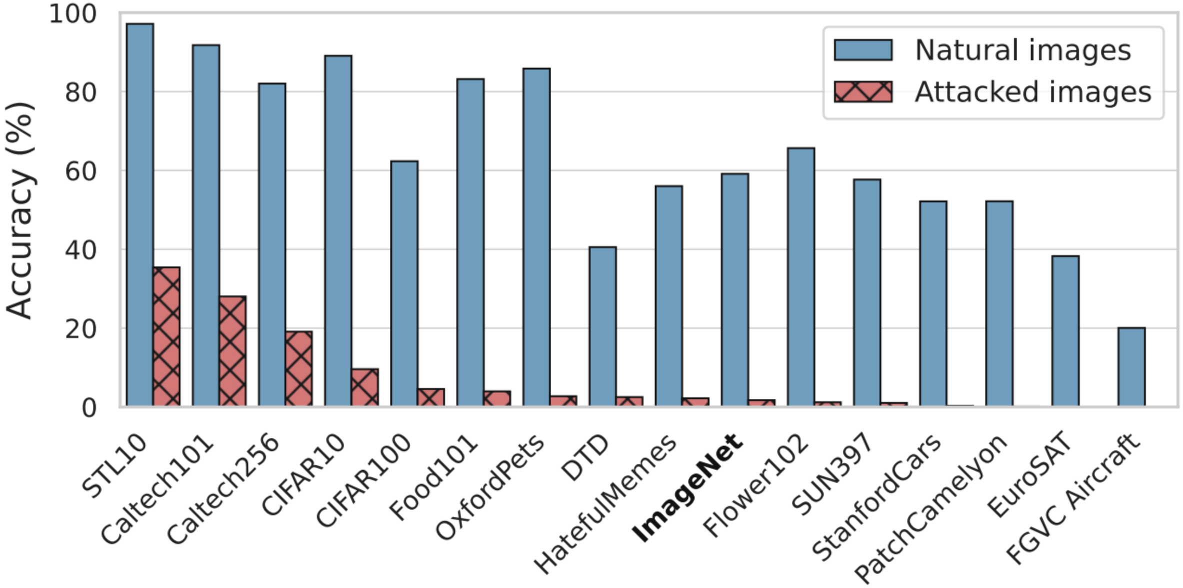
=

Adversarial Example
“chihuahua”

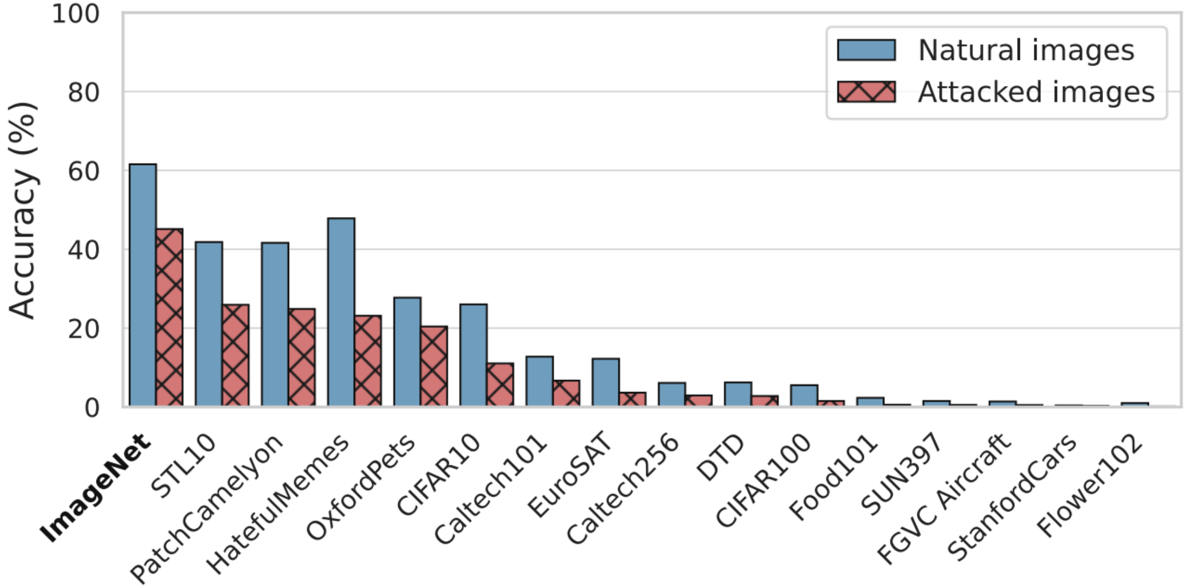
Adversarial Attack

NEURAL INFORMATION PROCESSING SYSTEMS

Imperceptible adversarial perturbations can significantly reduce CLIP’s performance on new tasks.



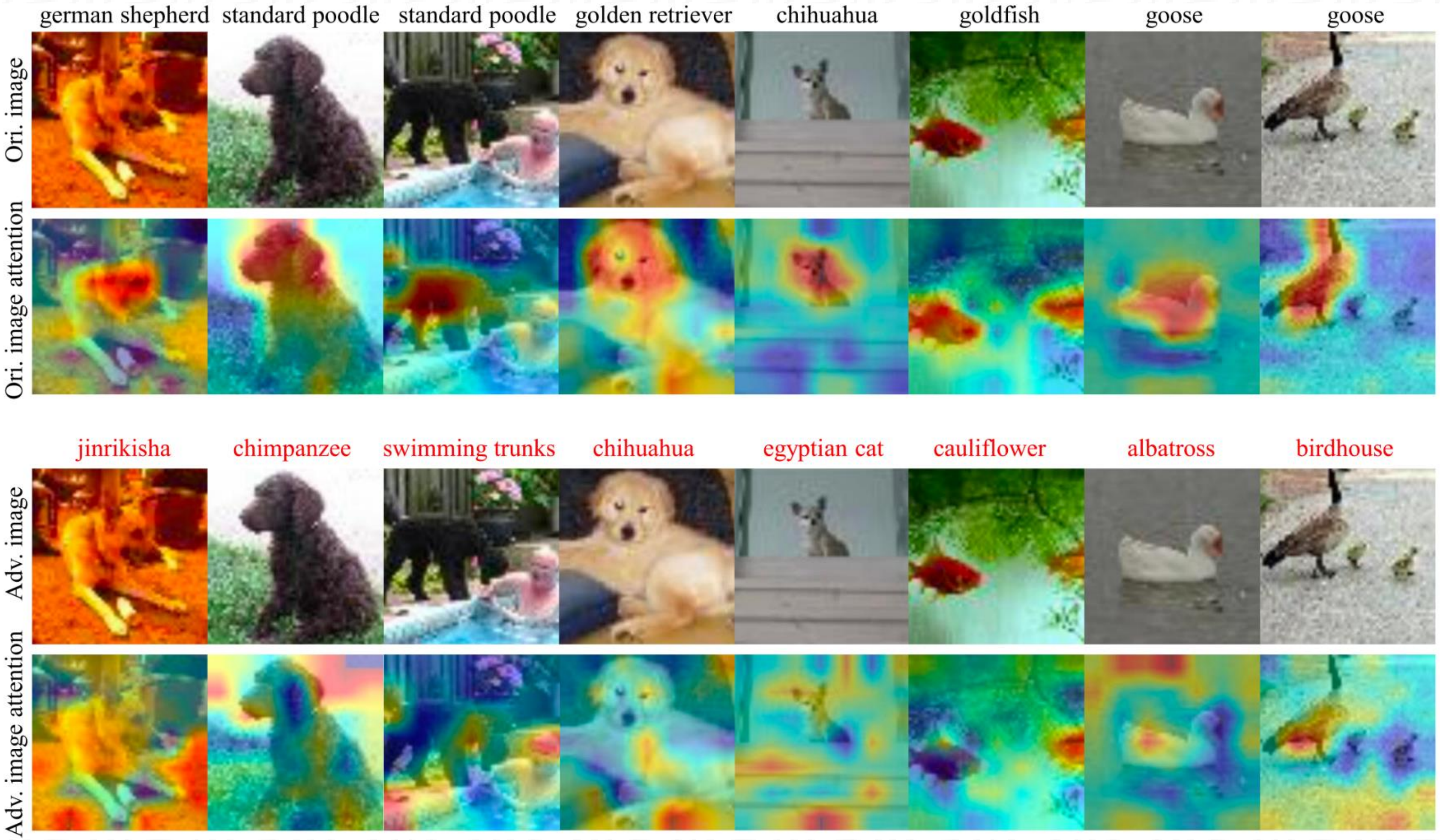
(a) CLIP



(b) Adversarially Finetuned CLIP

*Mao, C., Geng, S., Yang, J., Wang, X. and Vondrick, C., Understanding Zero-shot Adversarial Robustness for Large-Scale Models. In *The Eleventh International Conference on Learning Representations, 2023*.

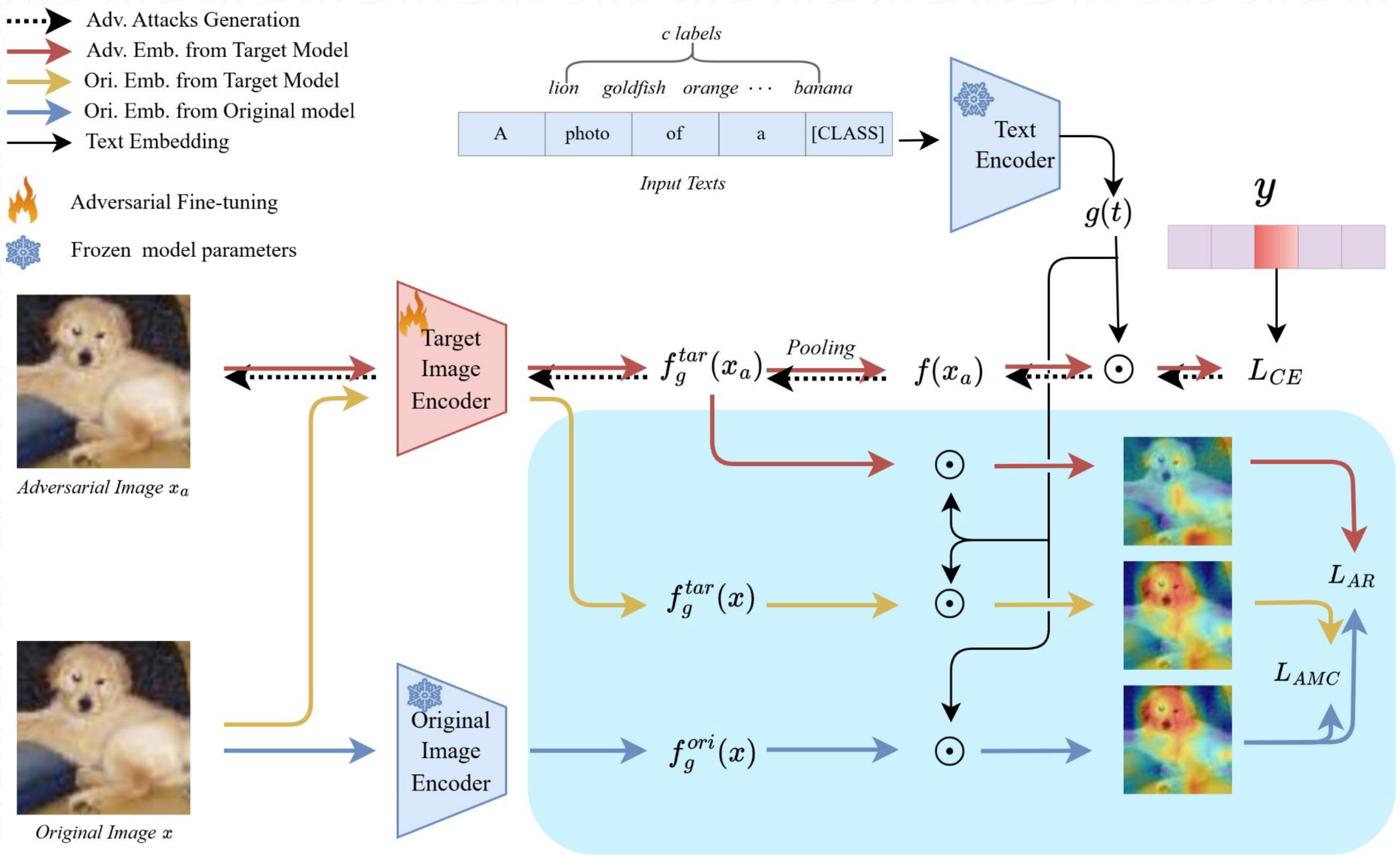
Interpretation of Adversarial Attacks



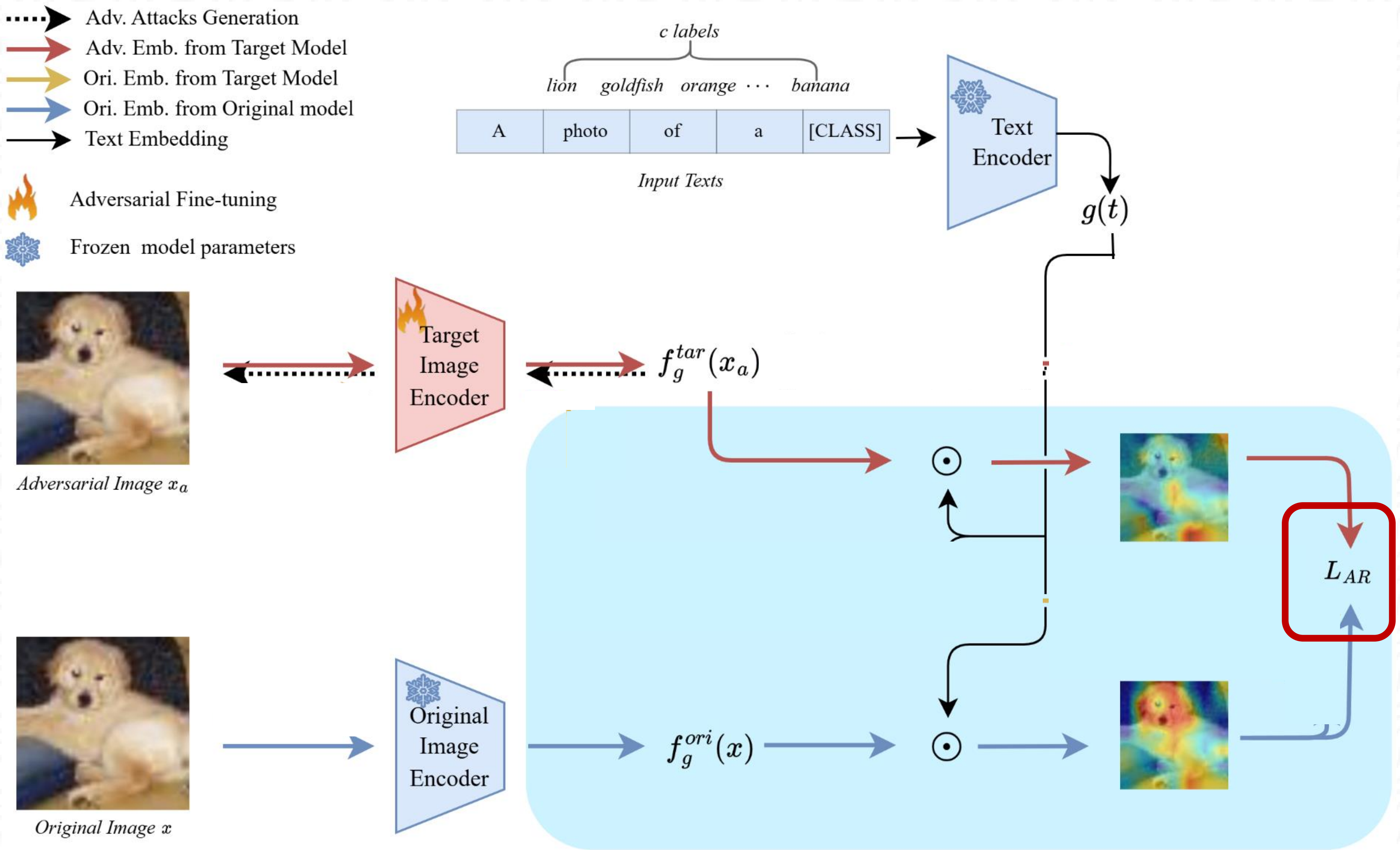
$$A(x) = f_g(x) \cdot g(t)^\top$$

We observe **a notable shift** in the **text-guided attention** of the adversarial example.

Text-Guided Attention for Zero-Shot Robustness



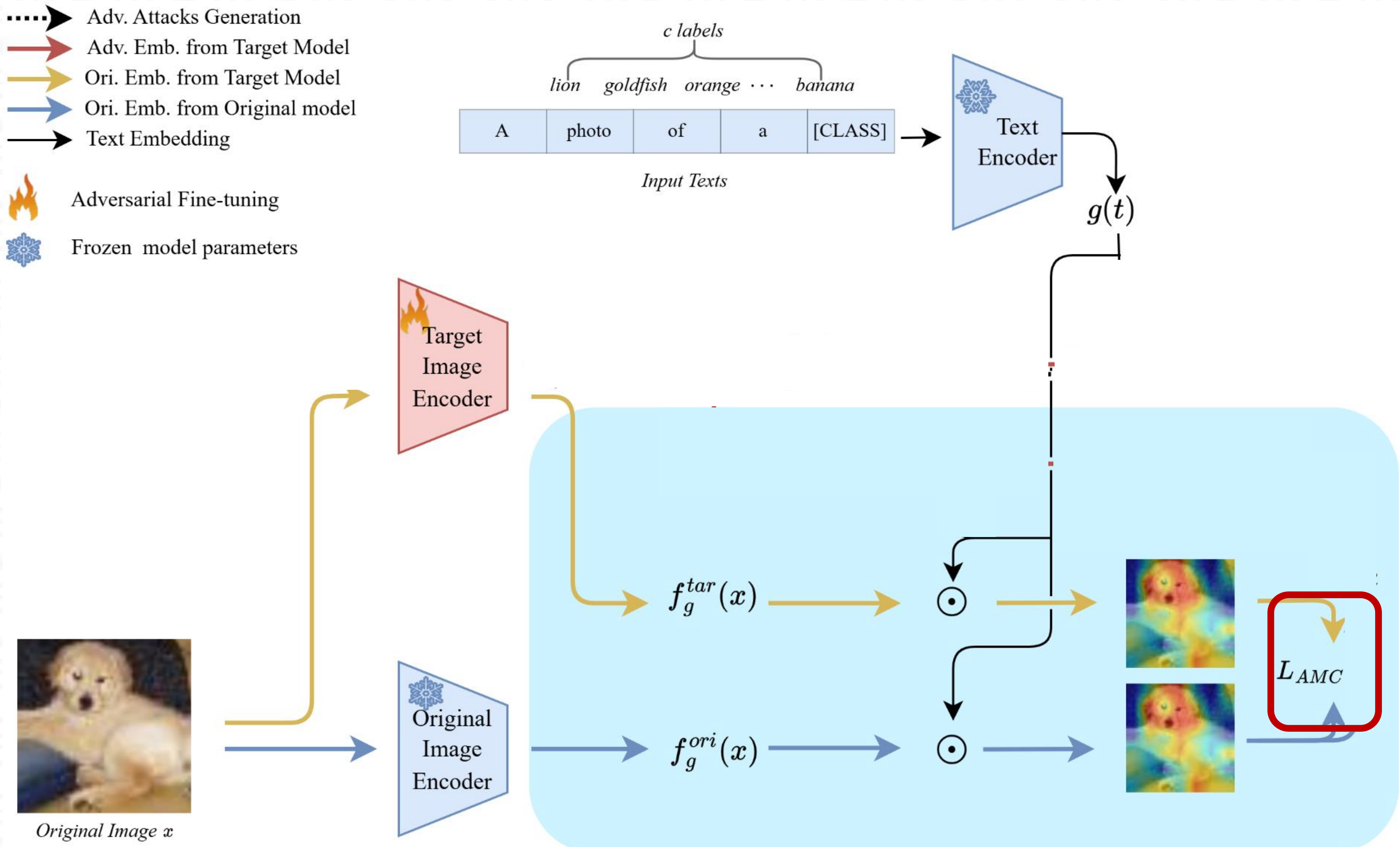
Text-Guided Attention for Zero-Shot Robustness



➢ **Attention Refinement Module:**

$$L_{AR} = \frac{1}{N} \cdot \sum_{i=0}^{N} \left\| A(x_a^i)_{tar} - A(x^i)_{ori} \right\|_2$$

Text-Guided Attention for Zero-Shot Robustness

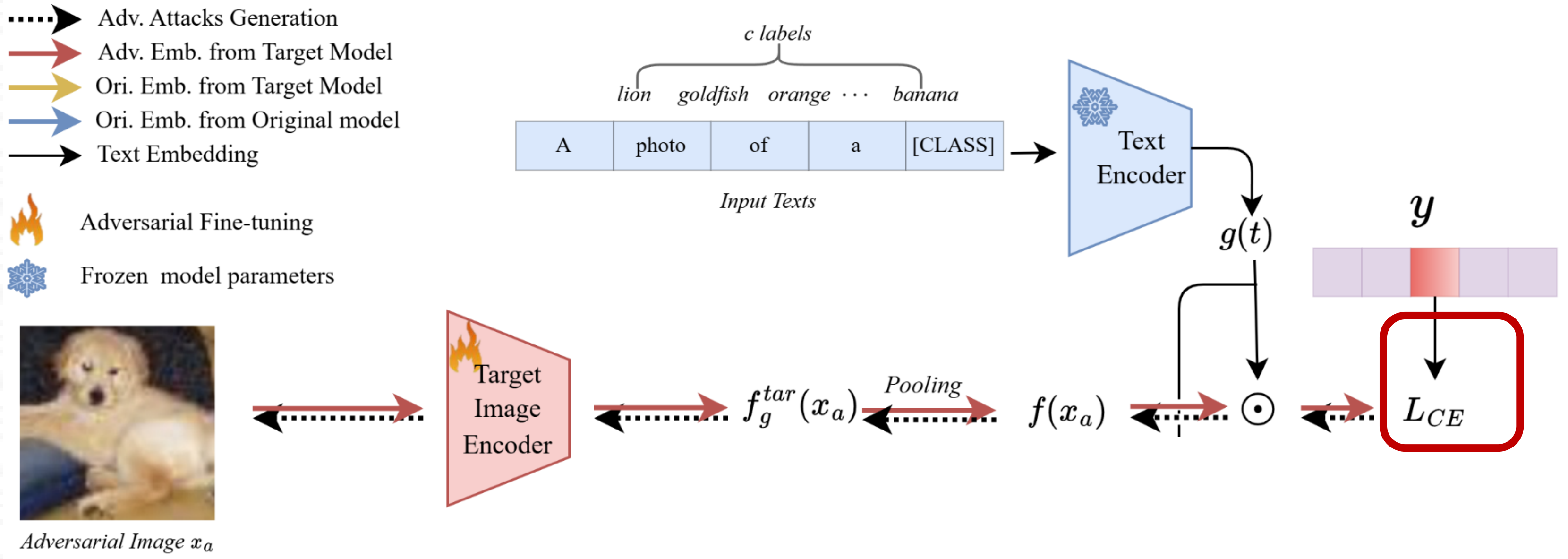


- **Attention-based Model Constraint Module:**

$$L_{AMC} = \frac{1}{N} \cdot \sum_{i=0}^{N} \left\| A(x^i)_{tar} - A(x^i)_{ori} \right\|_2$$

Text-Guided Attention for Zero-Shot Robustness

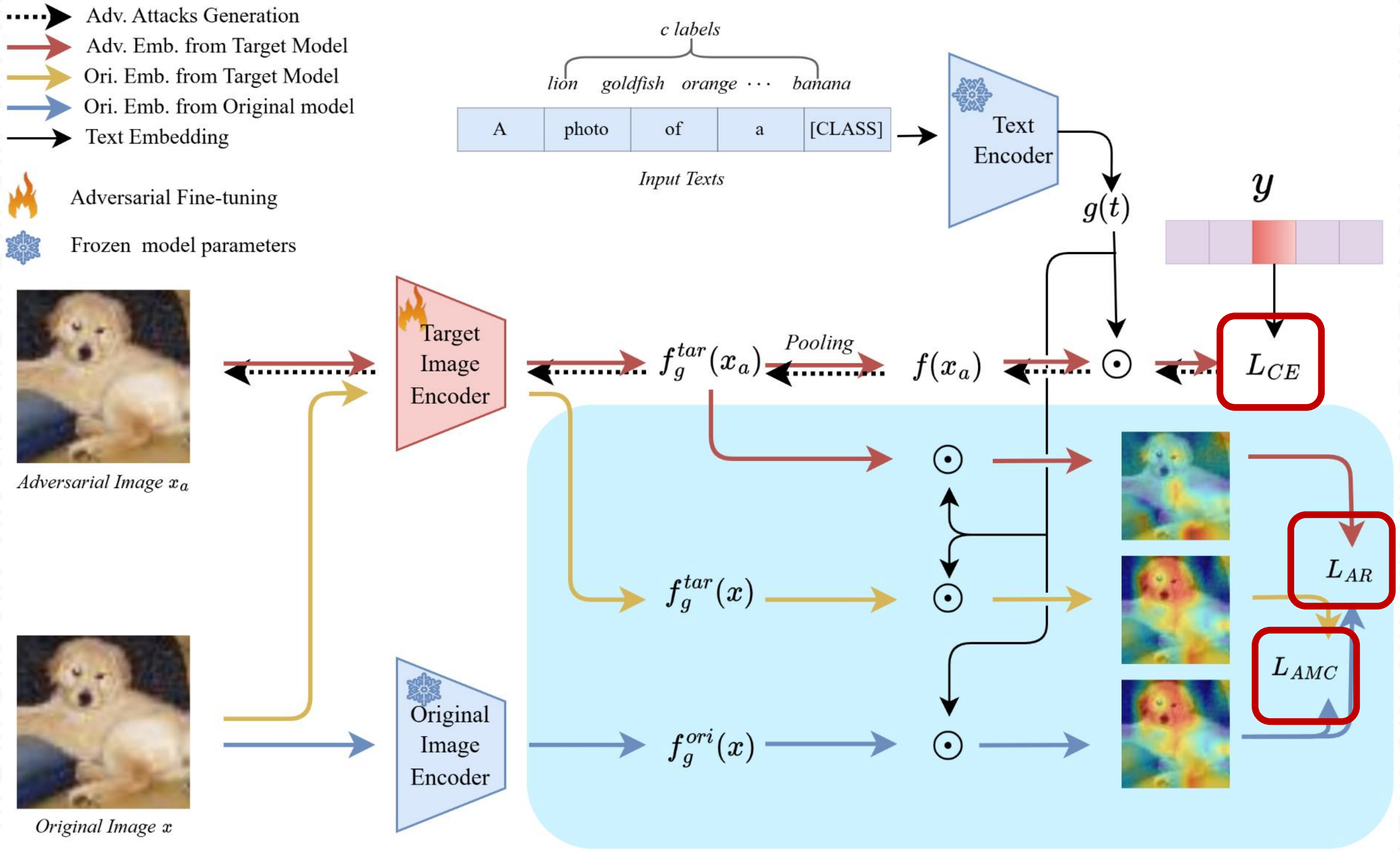
NEURAL INFORMATION PROCESSING SYSTEMS



- **Cross-Entropy Loss:**

$$L(x,t,y) = -\mathrm{E}_{i,j}[y_{ij} log \frac{\exp(\cos(f(x)_i, g(t)_j)/\tau)}{\sum_k \exp(\cos(f(x)_i, g(t)_k)/\tau)}]$$

Text-Guided Attention for Zero-Shot Robustness



Experiments: Main Results

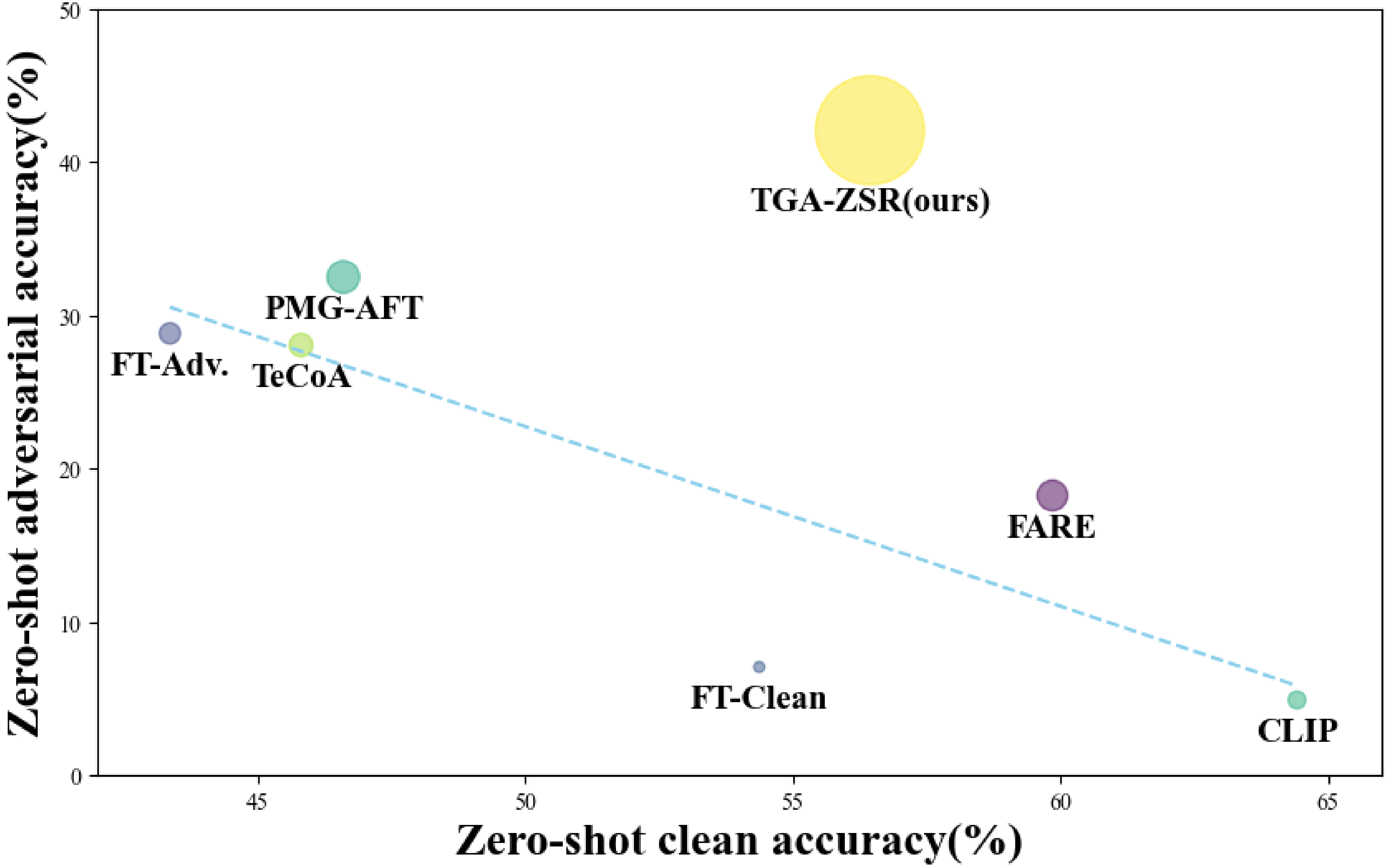
Table 1: Zero-shot robust accuracy on images attacked with 100 steps of PGD [36]. We performed several different methods on Tiny-ImageNet and evaluated across 16 datasets. The optimal accuracy is highlighted in **bold**, while the second-best accuracy is underlined. The values in parentheses represent the standard deviation.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	Oxfordpets	Flowers102	DTD	EuroSAT	FGVC-Aircraft	ImageNet	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP [48]	0.88	2.42	0.26	26.11	1.00	6.60	3.84	1.19	2.02	0.05	0.00	1.24	19.88	12.60	0.20	0.11	4.90
FT-Clean	13.55	19.92	4.94	40.00	0.82	0.64	2.40	0.68	2.66	0.05	0.03	1.08	14.95	9.69	0.09	1.32	7.05
FT-Adv.	51.59	38.58	21.28	69.55	17.60	12.55	34.97	19.92	15.90	11.95	1.83	17.26	50.73	40.18	8.42	**48.88**	28.83
TeCoA [38]	37.57	30.30	17.53	67.19	19.70	14.76	36.44	22.46	17.45	12.14	1.62	18.18	55.86	41.88	8.49	47.39	28.06
FARE[51]	23.88	21.25	10.72	59.59	8.30	10.97	24.56	15.48	10.96	0.14	0.84	10.54	45.96	34.35	4.38	10.17	18.25
PMG-AFT[59]	47.11	46.01	25.83	74.51	22.21	19.58	41.62	23.45	15.05	12.54	1.98	21.43	62.42	45.99	11.72	48.64	32.51
TGA-ZSR (ours)	**63.95** (± 0.11)	**61.45** (± 0.67)	**35.27** (± 0.07)	**84.22** (± 0.21)	**33.22** (± 0.39)	**33.97** (± 0.20)	**57.75** (± 0.76)	**34.55** (± 0.35)	**22.08** (± 0.16)	**14.27** (± 0.26)	**4.75** (± 0.27)	**28.74** (± 0.11)	**70.97** (± 0.42)	**60.06** (± 0.46)	**20.40** (± 0.68)	47.76 (± 0.35)	**42.09** (± 0.12)

Table 2: Zero-shot clean accuracy. We performed several different methods on Tiny-ImageNet and evaluated across 16 datasets. The values in parentheses represent the standard deviation.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	Oxfordpets	Flowers102	DTD	EuroSAT	FGVC-Aircraft	ImageNet	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP [48]	57.26	**88.06**	60.45	**97.04**	**57.26**	**83.89**	**87.41**	**65.47**	**40.69**	**42.59**	**20.25**	**59.15**	**85.34**	**81.73**	**52.02**	**52.09**	**64.42**
FT-Clean	**79.04**	84.55	54.25	93.78	46.80	47.10	80.98	46.43	30.32	24.39	9.30	44.40	78.69	70.81	31.15	47.89	54.37
FT-Adv.	73.83	68.96	39.69	86.89	33.37	27.74	60.10	33.45	23.14	16.49	4.86	32.06	67.41	57.72	18.11	49.91	43.36
TeCoA [38]	63.97	66.14	36.74	87.24	40.54	35.11	66.15	38.75	25.53	17.13	6.75	37.09	74.63	62.50	24.65	50.01	45.81
FARE[51]	77.54	87.58	**62.80**	94.33	49.91	70.02	81.47	57.10	36.33	22.69	14.19	51.78	84.04	77.50	44.35	46.07	59.85
PMG-AFT[59]	67.11	74.62	44.68	88.85	37.42	37.47	66.34	35.66	21.17	17.76	4.71	35.93	76.70	61.96	25.21	49.99	46.60
TGA-ZSR(ours)	75.72 (± 0.12)	86.46 (± 0.26)	56.52 (± 0.35)	93.48 (± 0.19)	51.99 (± 0.25)	57.59 (± 0.34)	77.32 (± 0.30)	48.08 (± 0.37)	29.06 (± 0.35)	24.24 (± 0.49)	11.93 (± 0.27)	48.04 (± 0.06)	80.70 (± 0.09)	74.74 (± 0.18)	36.62 (± 1.03)	49.58 (± 0.17)	56.44 (± 0.08)

Experiments: Main Results



Experiments: Comparison to Vision-based Attention

NEURAL INFORMATION PROCESSING SYSTEMS

Table 4: Comparison of vision-based attention and our text-guided attention. We evaluate the state-of-the-art method PMG-AFT alongside our pipeline, incorporating two different types of attention mechanisms on Tiny-ImageNet and evaluating performance across 16 datasets.

Test	Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	Oxfordpets	Flowers102	DTD	EuroSAT	FGVC-Aircraft	ImageNet	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
Robust	PMG-AFT[55]	47.11	46.01	25.83	74.51	22.21	19.58	41.62	23.45	15.05	12.54	1.98	21.43	62.42	45.99	11.72	**48.64**	32.51
	Vision-based	52.81	40.46	22.66	70.26	19.50	13.74	37.67	19.78	16.97	11.79	2.64	18.08	55.64	42.45	8.88	38.11	29.47
	TGA-ZSR (ours)	**63.97**	**61.82**	**35.25**	**83.99**	**32.78**	**34.13**	**56.91**	**34.20**	**21.92**	**14.20**	**4.44**	**28.62**	**70.53**	**59.70**	**21.15**	47.75	**41.96**
Clean	PMG-AFT[55]	67.11	74.62	44.68	88.85	37.42	37.47	66.34	35.66	21.17	17.76	4.71	35.93	76.70	61.96	25.21	49.99	46.60
	Vision-based	74.31	70.77	41.03	87.24	36.91	30.07	62.52	33.89	24.10	16.26	5.70	33.59	72.35	59.75	20.50	**51.29**	45.02
	TGA-ZSR (ours)	**76.85**	**86.23**	**56.55**	**93.28**	**51.71**	**57.72**	**77.08**	**48.32**	**29.15**	**23.99**	**12.03**	**48.10**	**80.82**	**74.58**	**37.72**	49.60	**56.48**

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments: More Attack

Table 3: Zero-shot robust accuracy on images attacked with ε of 1/255 of AutoAttack [7]. We performed several different methods on Tiny-ImageNet and evaluated on 16 datasets.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	Oxfordpets	Flowers102	DTD	EuroSAT	FGVC-Aircraft	ImageNet	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP [48]	0.02	0.01	0.08	0.03	0.04	0.01	0.00	0.03	0.16	0.12	0.06	0.04	0.43	0.10	0.11	0.22	0.09
FT-Clean	0.08	0.03	0.01	0.91	0.09	0.04	0.06	0.03	0.48	0.02	0.03	0.12	1.38	0.66	0.03	0.03	0.25
FT-Adv.	**50.48**	37.55	20.39	69.14	16.25	11.23	33.91	18.54	**19.95**	11.59	1.65	16.21	49.90	39.24	7.57	**48.84**	28.28
TeCoA [38]	35.03	28.18	16.09	66.08	17.41	13.05	34.81	20.80	15.37	11.40	1.32	16.32	54.54	40.15	7.15	47.12	26.55
FARE [51]	28.59	23.37	13.58	60.70	9.72	13.88	27.72	15.48	9.15	0.25	0.87	12.07	47.45	36.68	6.77	10.23	19.78
PMG-AFT [59]	44.26	**44.12**	**23.66**	**73.90**	19.63	**17.25**	39.25	20.87	13.72	**11.99**	1.68	19.17	**60.57**	44.25	9.59	48.53	30.78
TGA-ZSR (ours)	49.45	40.53	22.38	72.06	**20.36**	15.58	**40.31**	**21.43**	17.13	11.19	**2.64**	**19.28**	57.16	**45.68**	**10.47**	48.03	**30.86**

Table 4: Zero-shot robust accuracy across 16 datasets with CW attack [4]. The optimal accuracy is highlighted in **bold**.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	Oxfordpets	Flowers102	DTD	EuroSAT	FGVC-Aircraft	ImageNet	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP [48]	0.21	0.36	0.10	10.59	1.16	0.82	1.23	1.09	2.18	0.01	0.00	1.14	13.50	7.36	2.36	0.07	3.64
PMG-AFT[59]	44.59	44.86	24.15	74.11	19.99	17.33	39.88	20.95	13.51	12.09	1.47	19.51	60.99	44.46	10.57	**48.59**	31.07
TGA-ZSR(ours)	**63.85**	**60.50**	**34.62**	**84.11**	**22.03**	**33.28**	**58.33**	**32.95**	**21.22**	**13.89**	**4.56**	**20.42**	**70.34**	**59.73**	**20.20**	48.02	**40.50**

Experiments: Comparison to Computational Overhead

NEURAL INFORMATION PROCESSING SYSTEMS

Table 8: Comparison of memory usage, training time, and test time.

Methods	Train memory usage	Train time (per epoch / batch)	Test time (per batch)
CLIP [48]	0Mb	0s / 0s	21s
TeCoA [38]	12873Mb	512s / 0.65s	21s
PMG-AFT[59]	18449Mb	828s / 1.06s	21s
TGA-ZSR (ours)	$\boxed{21227Mb}$	$\boxed{885s / 1.13s}$	21s

Conclusions

In this paper, we discovered that ***adversarial attacks lead shift of text-guided attention***. Building on this observation, we introduce a text-guided approach, ***TGA-ZSR***, which incorporates two key components to preform adversarial fine-tuning and constrain the model. ***This strategy prevents model drift while enhancing model robustness.***

Thank You!

luyu@email.tjut.edu.cn
zshy@stud.tjut.edu.cn
csxu@nlpr.ia.ac.cn

https://github.com/zhyblue424/TGA-ZSR