

# Quality Text, Robust Vision: The Role of Language in Enhancing Visual Robustness of Vision-Language Models

**Futa Waseda**<sup>1,2</sup>, Saku Sugawara<sup>2</sup>, Isao Echizen<sup>2</sup>

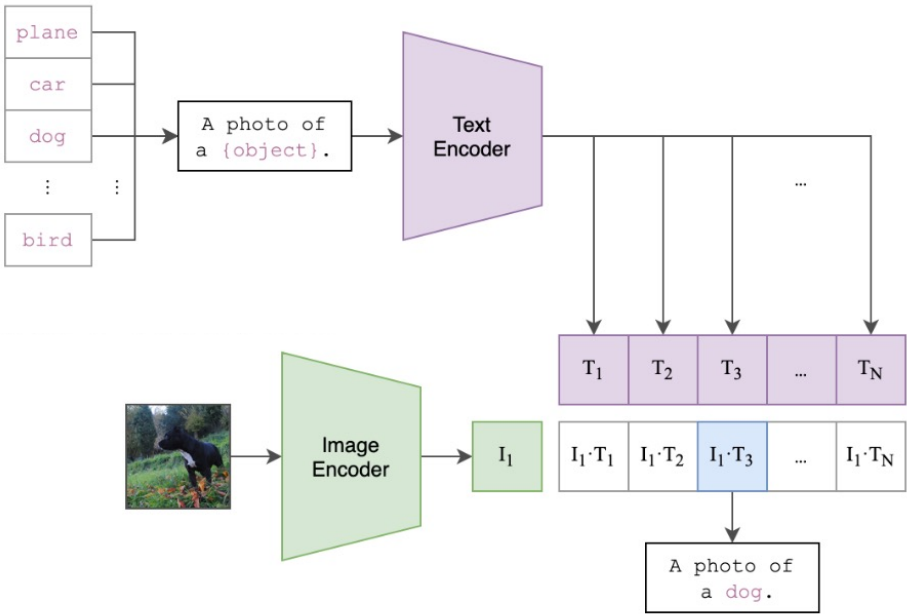
<sup>1</sup> The University of Tokyo,    <sup>2</sup> National Institute of Informatics

Correspondence to: Futa Waseda <[futa-waseda@g.ecc.u-tokyo.ac.jp](mailto:futa-waseda@g.ecc.u-tokyo.ac.jp)>

# Problem Settings: Zero-shot Adversarial Robustness 2

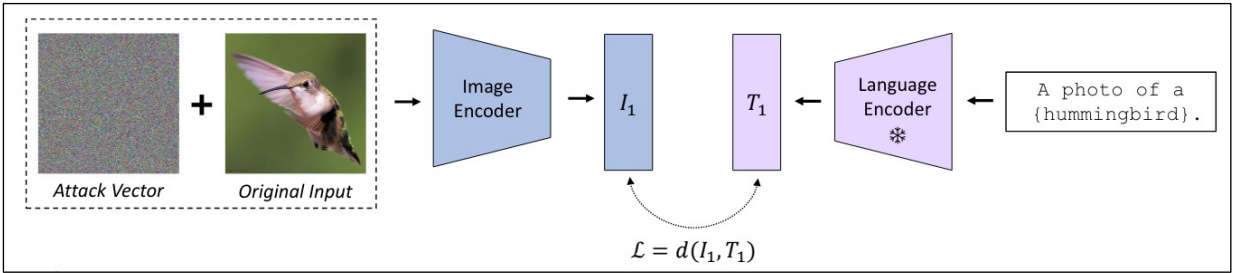
CLIP enables zero-shot image recognition; However, it is vulnerable to adversarial attacks

## CLIP: Zero-shot image classification



## Robust Fine-tuning for CLIP

TeCoA [ICLR2023]: **Adversarial Fine-tuning** of CLIP



## Robust Fine-tuning on ImageNet

| Method          | ImageNet | DATASETS |        |          |       |              |         |           |           |       |         |      |           |       |            |            |
|-----------------|----------|----------|--------|----------|-------|--------------|---------|-----------|-----------|-------|---------|------|-----------|-------|------------|------------|
|                 |          | CIFAR10  | STL-10 | CIFAR100 | SUN   | StanfordCars | Food101 | OxfordPet | Flower102 | DTD   | EUROSAT | FGVC | HateMemes | PCAM  | Caltech101 | Caltech256 |
| +FT(Adv.)       | 45.12    | 11.04    | 25.90  | 1.51     | 0.52  | 0.25         | 0.56    | 20.40     | 0.11      | 2.77  | 3.62    | 0.48 | 23.12     | 24.86 | 6.68       | 2.92       |
| +FT (CoAdv.)    | 33.41    | 2.51     | 5.92   | 0.40     | 0.01  | 0.14         | 0.09    | 0.19      | 0.47      | 0.21  | 6.62    | 0.42 | 30.2      | 44.7  | 0.06       | 0.08       |
| +FT (ImgCoAdv.) | 0.03     | 4.00     | 3.76   | 0.37     | 0.12  | 0.05         | 0.17    | 0.57      | 0.21      | 0.96  | 4.13    | 0.18 | 32.93     | 26.41 | 0.33       | 0.18       |
| FT (TeCoA)      | 41.88    | 59.28    | 83.45  | 34.13    | 29.81 | 13.37        | 27.99   | 62.61     | 30.69     | 22.88 | 15.18   | 5.10 | 31.97     | 23.87 | 69.07      | 59.54      |
|                 |          | Average  |        |          |       |              |         |           |           |       |         |      |           |       |            |            |
|                 |          | 10.62    |        |          |       |              |         |           |           |       |         |      |           |       |            |            |

Evaluated on diverse zero-shot (unseen) datasets

# Issue: Existing Defenses Do Not Use “Language”

3

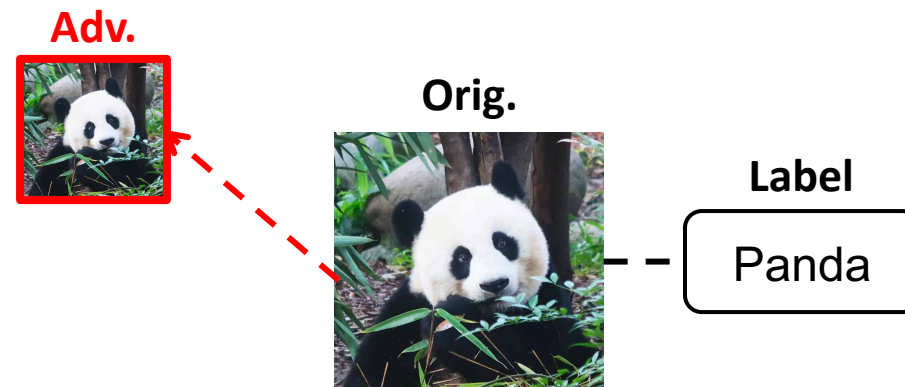
Baselines **use only class labels**, or **does not use language (unsupervised)** to create adversarial examples

## Supervised Adv. Finetuning (e.g, TeCoA)



Only deviates from class labels  
=> **Overfits to training data**

## Unsupervise Adv. Fine-tuning (e.g., FARE)



Deviates from original image  
=> **Fail to corrupt semantic features**

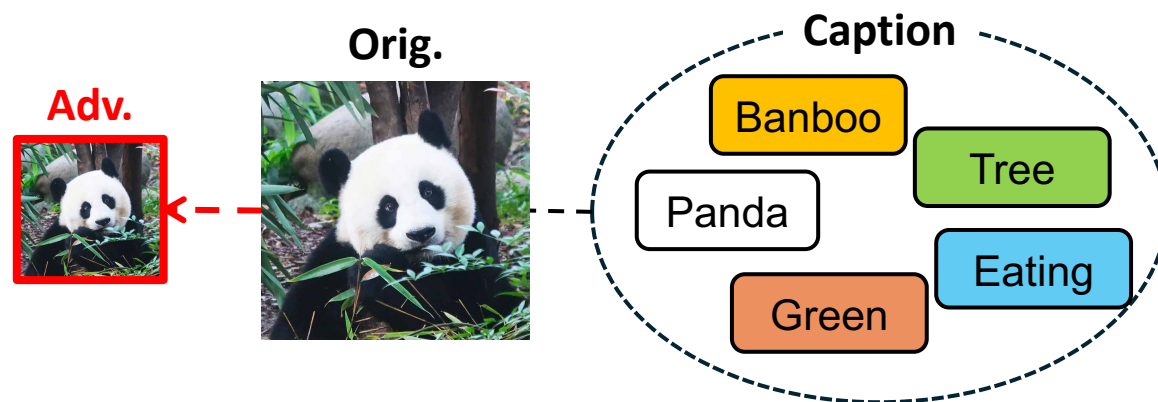
# Method: Use Rich Texts for Semantic Guidance

4

## Quality Text-guided Adversarial Fine-Tuning (QT-AFT)

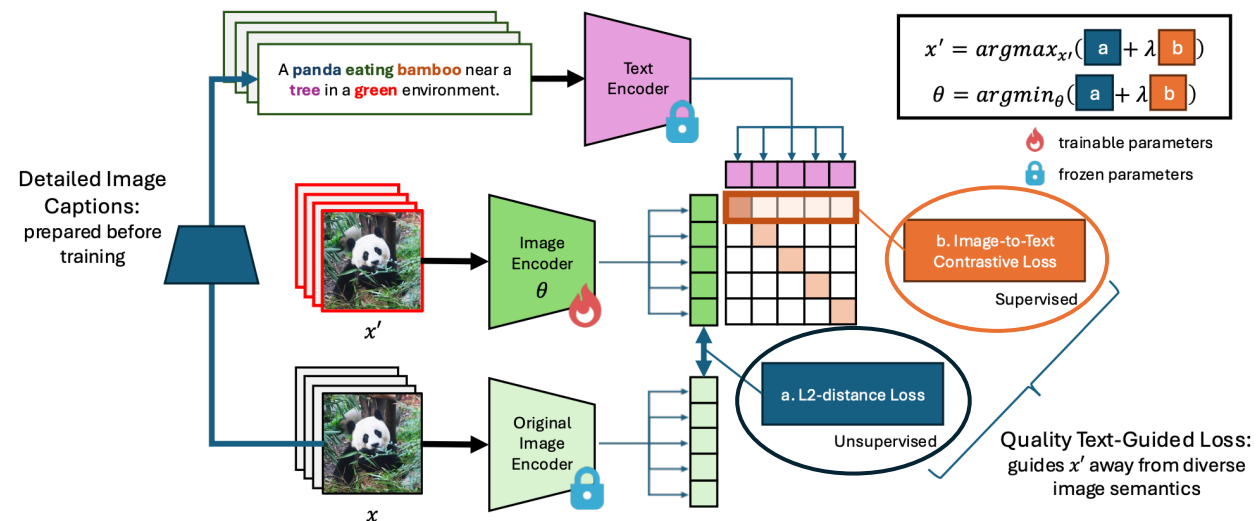
Use detailed captions (~50 words) as textual supervision during AT

### Idea



=> Improves zero-shot robustness

### Training Details



1. Prepare: Generate captions for all images
2. Adversarial Training:
  1. Use captions for supervision loss
  2. Combine with unsupervised loss

## QT-AFT achieved SOTA clean/robust accuracy

|        |                      | Zero-shot datasets |             |             |             |             |             |             |             |             |             |             |              |             |             |             |             |             |                |
|--------|----------------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|----------------|
| Method |                      | ImageNet           | ImageNet-S  | ImageNet-R  | CIFAR-10    | CIFAR-100   | STL-10      | Caltech101  | Caltech256  | OxfordPets  | Flowers102  | FGVC        | StanfordCars | SUN397      | Food101     | EuroSAT     | DTD         | PCAM        | Avg. Zero-shot |
| Clean  | PMG-AFT              | 55.6               | 31.7        | 50.9        | <u>76.6</u> | 45.9        | 92.5        | 77.7        | 67.5        | 67.1        | 9.9         | 2.9         | 8.6          | 34.9        | 27.9        | <b>23.5</b> | 24.8        | 48.0        | 43.1           |
|        | TGA-ZSR              | <b>83.4</b>        | <b>42.3</b> | 52.4        | <b>88.7</b> | <b>58.7</b> | <b>96.6</b> | 15.1        | <b>83.9</b> | 67.3        | <b>48.7</b> | 8.5         | 36.7         | <b>64.8</b> | <b>77.0</b> | 0.2         | <b>29.6</b> | 47.5        | <u>51.1</u>    |
|        | TeCoA                | <u>63.3</u>        | 31.8        | 51.9        | 75.2        | 39.1        | 91.7        | 74.7        | 66.4        | 71.8        | 19.5        | 6.9         | 12.6         | 34.7        | 20.9        | 17.0        | 21.4        | <b>57.9</b> | 43.3           |
|        | FARE                 | 50.6               | 35.6        | <b>57.0</b> | 64.5        | 47.3        | 91.8        | <u>80.5</u> | <u>74.4</u> | <u>76.4</u> | <u>39.1</u> | <b>13.5</b> | <u>39.5</u>  | 43.8        | <u>44.3</u> | <u>21.9</u> | 27.0        | 48.0        | 50.3           |
|        | <b>(ours) QT-AFT</b> | 51.9               | <u>38.5</u> | <u>56.9</u> | 70.9        | <u>48.6</u> | <u>95.8</u> | <b>81.9</b> | 73.4        | <b>80.7</b> | 30.6        | <u>12.5</u> | <b>40.1</b>  | <u>51.7</u> | 44.2        | 19.2        | <u>29.2</u> | <u>51.1</u> | <b>51.6</b>    |
| Adv.   | PMG-AFT              | 30.1               | <u>14.8</u> | 24.9        | <b>36.7</b> | <u>17.7</u> | <b>70.3</b> | <u>55.8</u> | 35.8        | <b>39.6</b> | 3.0         | 0.3         | 1.1          | 10.1        | 5.6         | 3.1         | 10.4        | <u>47.7</u> | <u>23.6</u>    |
|        | TGA-ZSR              | <u>31.1</u>        | 3.2         | 10.9        | 24.5        | 8.5         | 54.4        | 10.3        | 30.0        | 13.5        | 2.8         | 0.0         | 0.1          | 0.5         | 5.0         | 0.0         | 0.4         | 0.0         | 10.3           |
|        | TeCoA                | <b>32.8</b>        | 14.3        | <b>25.2</b> | 32.2        | 16.8        | 68.6        | 49.0        | <u>36.3</u> | <u>39.2</u> | 5.8         | 1.2         | 2.6          | <u>10.5</u> | 6.0         | <u>9.8</u>  | 10.2        | 20.5        | 21.8           |
|        | FARE                 | 20.0               | 14.0        | 20.9        | 30.7        | 15.2        | 62.6        | 53.0        | 35.8        | 30.4        | 8.6         | 1.8         | <u>2.8</u>   | 9.8         | 7.4         | 3.6         | 13.2        | <b>48.0</b> | 22.4           |
|        | <b>(ours) QT-AFT</b> | 19.6               | <b>17.6</b> | 25.2        | <u>33.2</u> | <b>20.9</b> | <u>69.0</u> | <b>58.9</b> | <b>40.6</b> | 36.5        | <b>9.7</b>  | <b>2.3</b>  | <b>5.8</b>   | <b>14.2</b> | <b>7.7</b>  | <b>12.6</b> | <b>14.6</b> | 44.1        | <b>25.8</b>    |

- Pointed out that **existing defenses for robust vision** overlook the **importance of language**
- Proposed QT-AFT
  - **High-quality language**-guided adversarial fine-tuning
  - **SOTA** in both clean and robust zero-shot performance
- Opens a novel direction: **leveraging multimodality to enhance vision robustness**  
—an innovative perspective beyond unimodal paradigms