Language as an Adversary for Vision-Language Models

Adham Ibrahim Sanoojan Baliah Jameel Hassan
{adham.ibrahim, sanoojan.baliah, jameel.hassan}@mbzuai.ac.ae

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Problem Statement

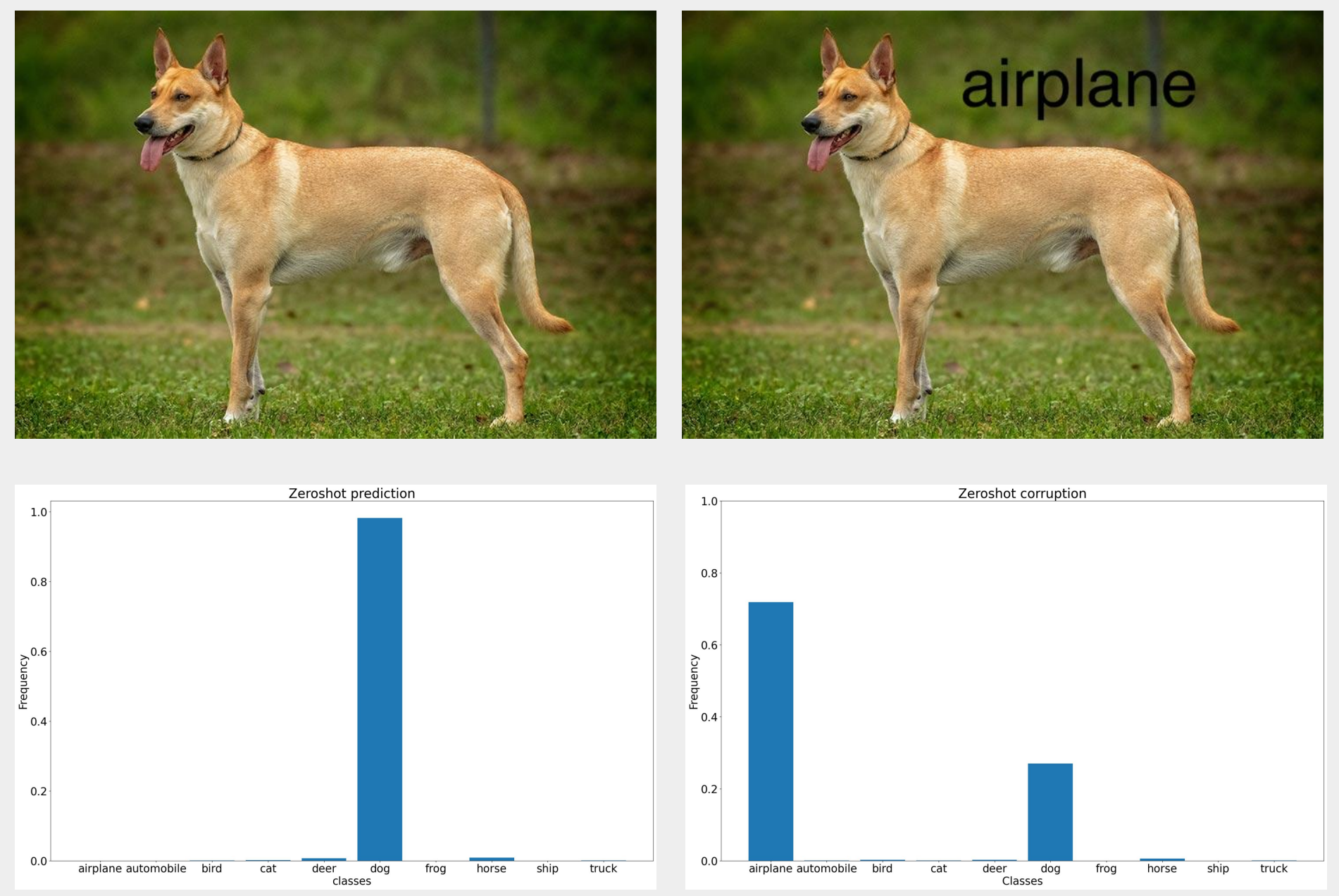
Problem: The vulnerability of large Vision and Language model CLIP for targeted attacks using language

Why CLIP? Advent of CLIP as a foundation model for many downstream task

Why targeted attack? A more challenging task while creating severe vulnerability to downstream tasks as the attacker has full control of the targeted label.

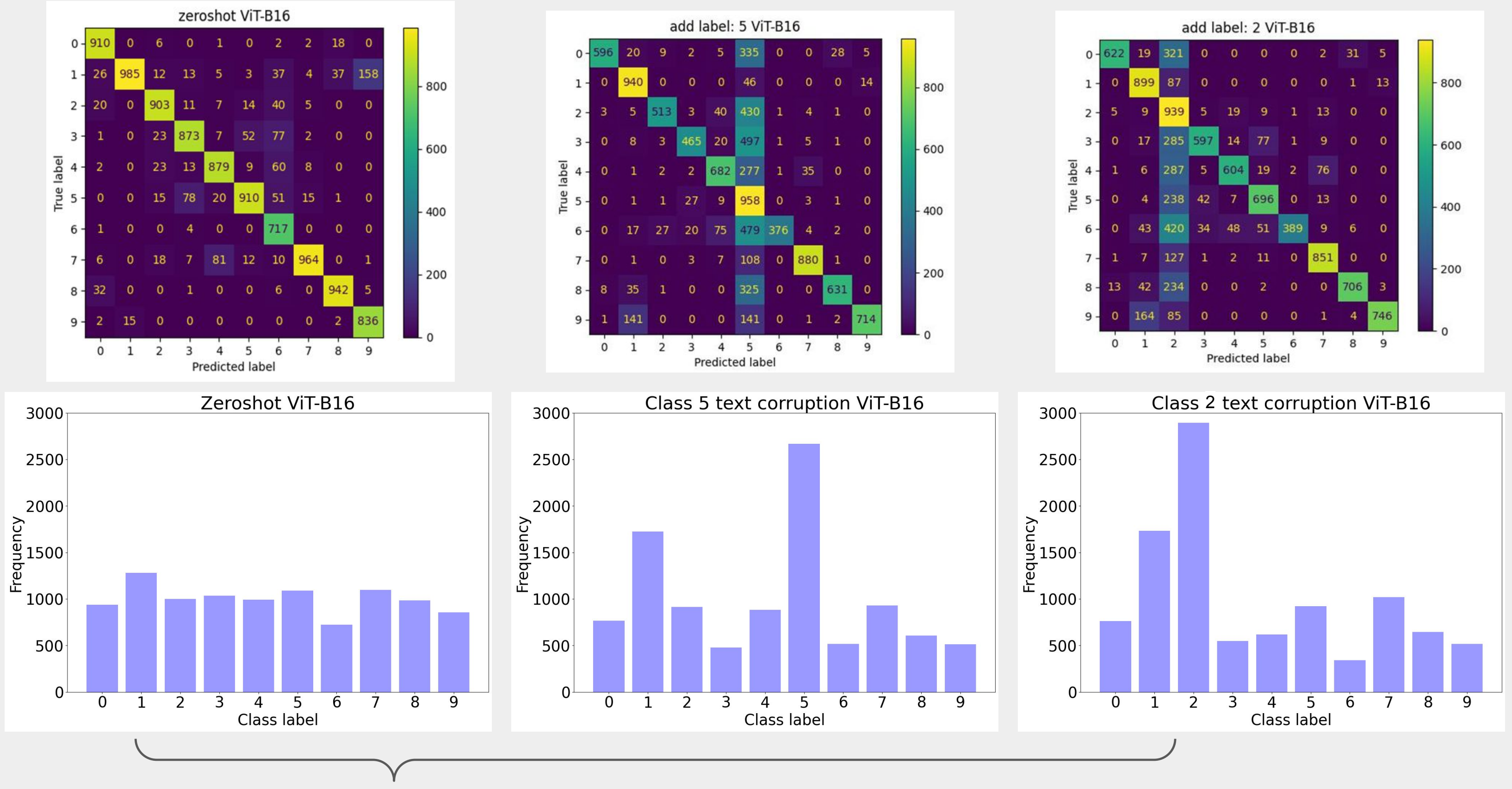
Goal: A whitebox targeted attack on CLIP model

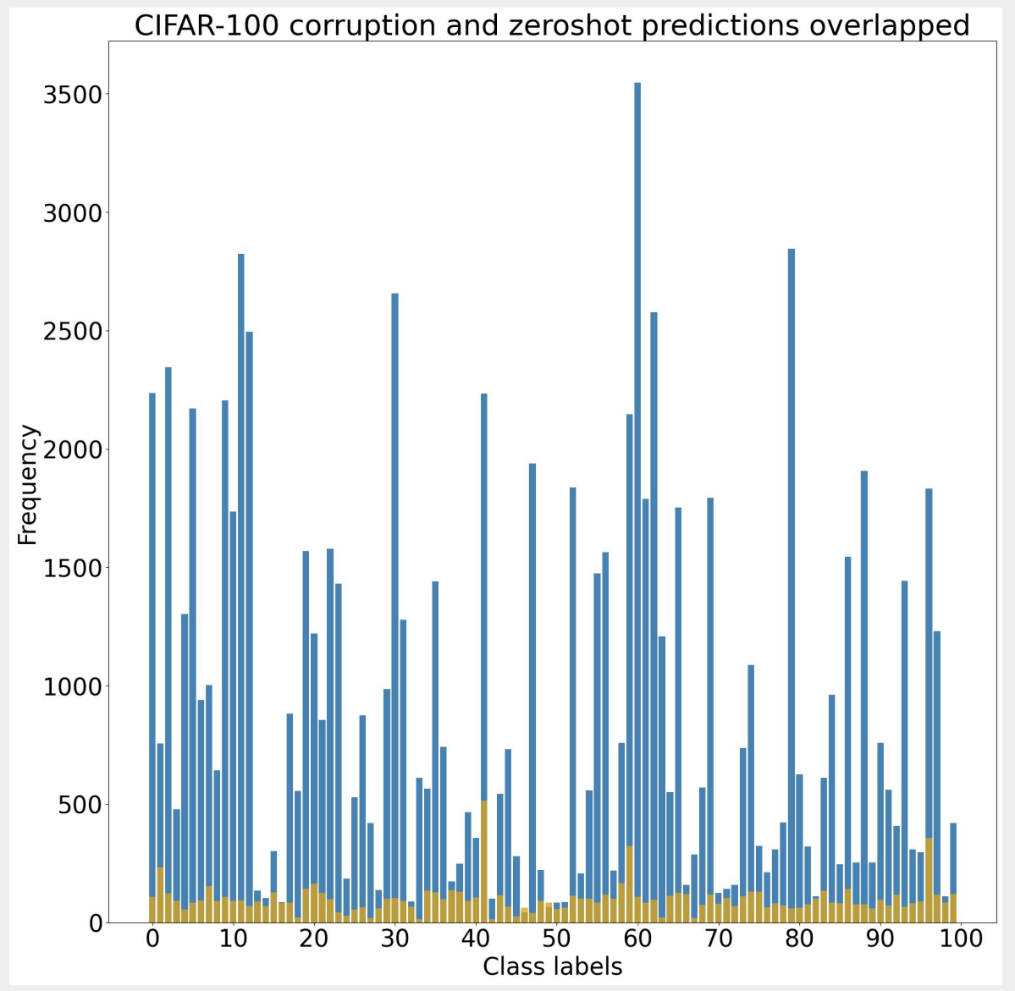
Motivation



The joint training of vision and language modalities causes the CLIP model to attend to text in images.

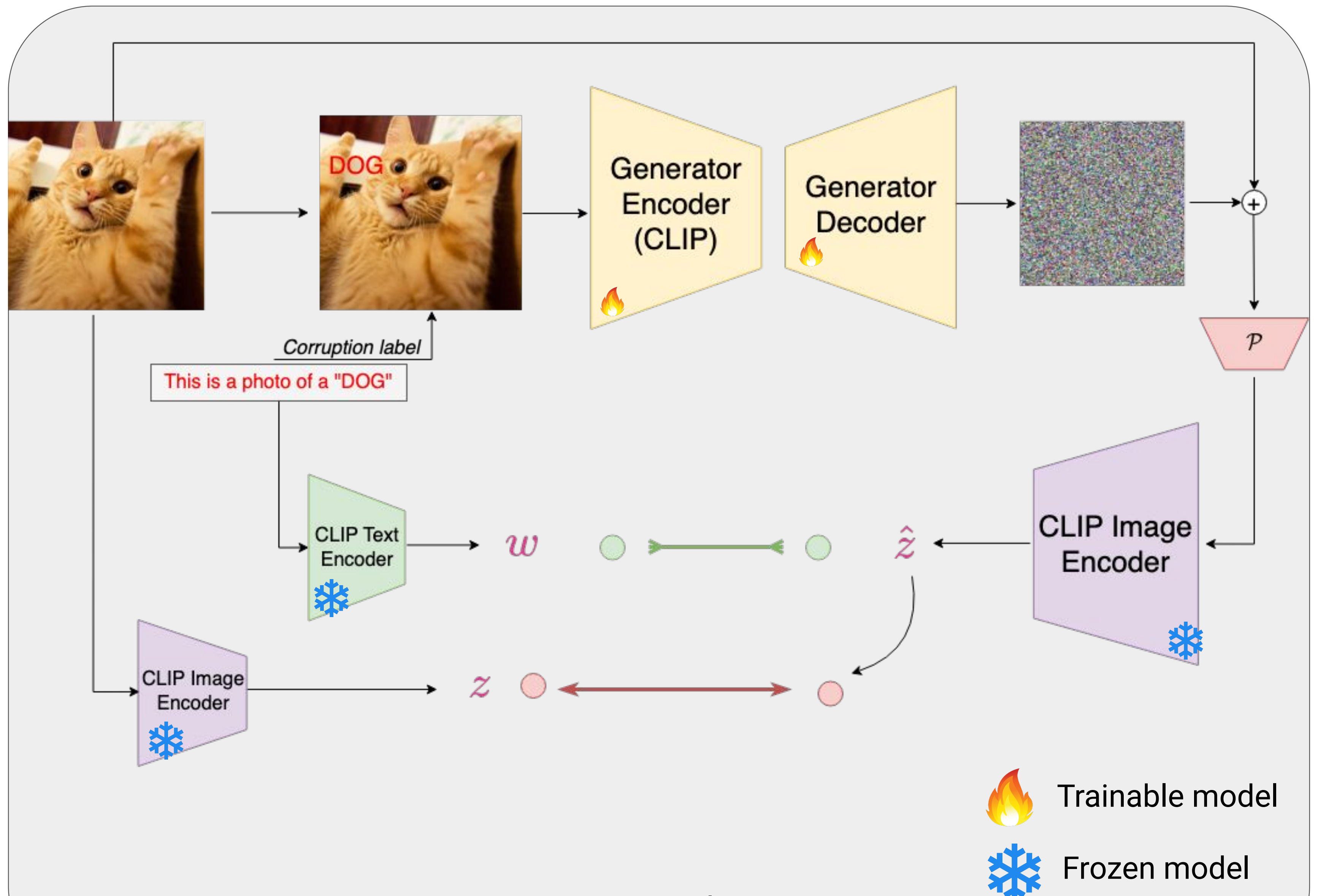
Can the text, biases the prediction more?
If so, can this be a reason for the downfall of vision and language models?





Zero shot predictions of CLIP model and the aggregation of the predictions for each class when that specific class is used as corruption on the CIFAR-100 dataset.

Methodology



Training pipeline

Design Factors

Generator model: Incorporates a ResNet-50 architecture encoder decoder. The encoder is initialized with CLIP pretrained ResNet-50 model. This enables the text conditioning to better be captured in the generator.

Projection Layer: The perceptual budget for adversarial attack is maintained using the projection layer with an L-∞ norm [1]. A perceptual budget of 0.1 is used.

Loss: A contrastive loss based on cosine similarity is used.
The adversarial image embedding similarity to the corruption text word embedding is maximized, while the similarity to the original image embedding is minimized.

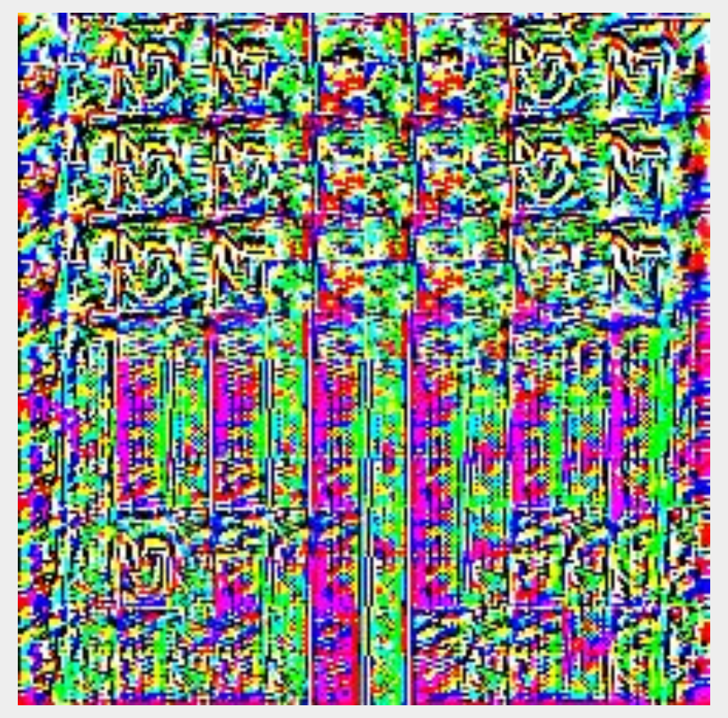
$$\mathcal{L}_{pos} = 1 - \left(\langle \hat{z}, \hat{w} \rangle\right)$$

$$\mathcal{L}_{neg} = \max(0, \langle (z, \hat{z}) \rangle - \psi) \quad \text{where } \psi \text{ is the margin}$$

$$\mathcal{L} = \mathcal{L}_{pos} + \mathcal{L}_{neg}$$

Original image

+



Adversarial Noise

=

Adversarial Image

Experiments & Results

Results with our attack strategy

Method	Dataset	Top1	Top 5	Attack 1	Attack 5
CLIP Zero shot	CIFAR-10	89.16	99.08	–	–
	CIFAR-100	64.40	86.65	–	–
	Caltech-101	83.21	96.06	–	–
Language written attack (Ours)	CIFAR-10	10.61	52.2	82.67	97.81
	CIFAR-100	2.63	8.81	1.13	5.11
	Caltech-101	46.09	70.04	1.26	5.74

Transferability to other datasets

Transfer	Top1	Top 5	Attack 1	Attack 5
CIFAR-10 to CIFAR-100	0.96	6.39	80.06	94.82
CIFAR-10 to Caltech-101	71.92	95.60	1.03	8.86

Ablation on Encoder type and learning rate

Generator Encoder	Lr	Top1	Top 5	Attack 1	Attack 5
Native ResNet50	1e-3	57.45	50.07	1.03	8.86
Unfrozen CLIP ResNet50	1e-5	11.12	51.68	70.68	94.82
	1e-3	10.61	52.2	82.67	97.81

Contributions and Discussions

- Language acts as an adversary to the CLIP vision-language model when it is naively written in the image itself.
- An architecture to map the corrupted image to an actual adversarial image which is indistinguishable as adversary to human.
- Extensive experiments shows this approach is able to create a text based targeted adversarial attack especially on CIFAR-10 for CLIP.

This generated adversarial images reduce the Top-1 accuracy of the CLIP ViT B/16 in all datasets. Further, the generator trained on CIFAR-10 attacks well on CIFAR-100 with the same target class space, but underperforms on Caltech-101, possibly due to the input pixel level information variation. This is an avenue that can be explored with more robust generators.

[1] Abhishek Aich, Calvin-Khang Ta, Akash A Gupta, Chengyu Song, S. Krishnamurthy, M. S. Asif, & Amit R. Chowdhury. GAMA: Generative adversarial multi-object scene attacks. In Thirty-Sixth Conference on NIPS, 2022