On the Robustness of Large Multimodal Models Against Image Adversarial Attacks

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, Ser-Nam Lim

Introduction

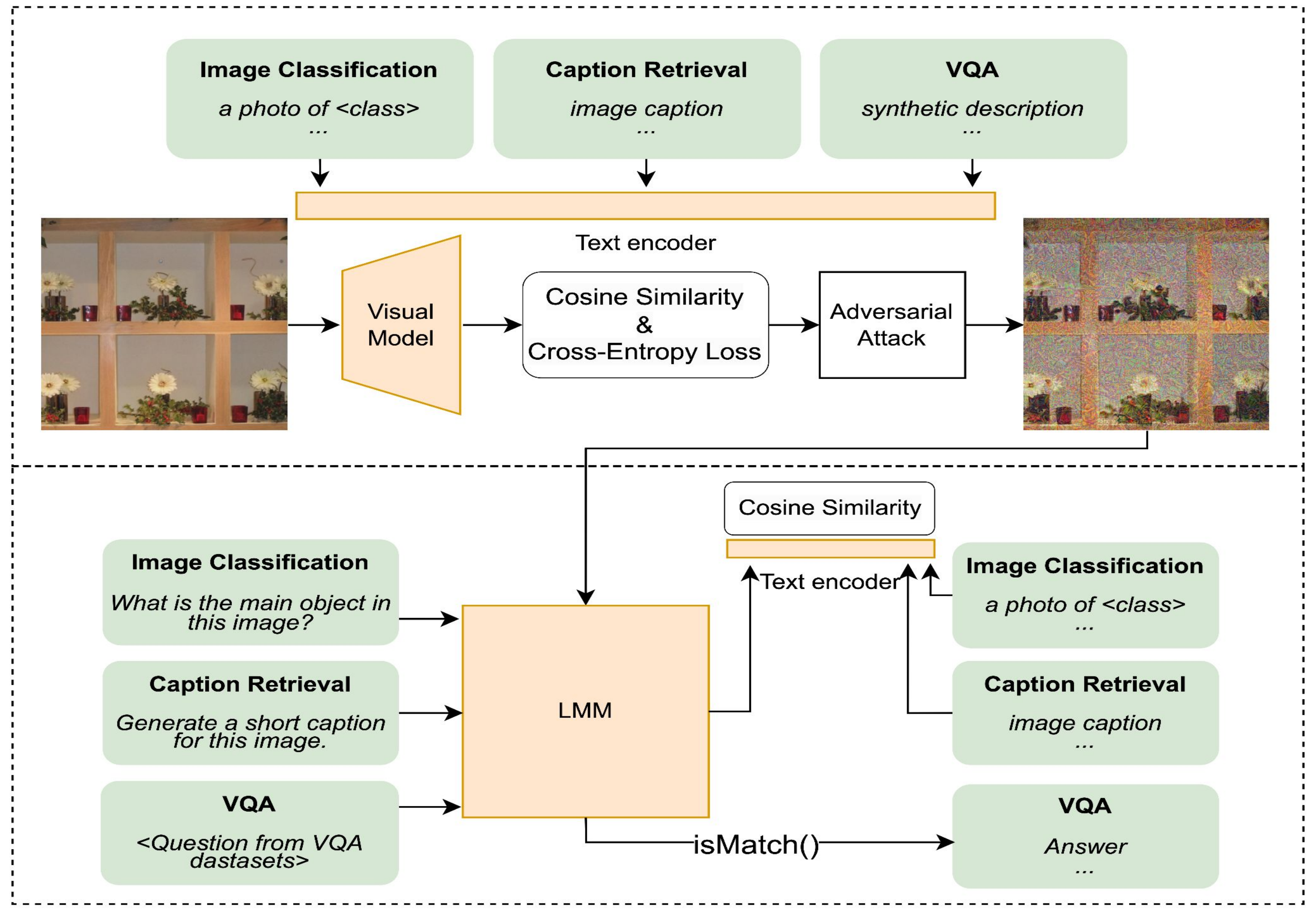
Overview:

A comprehensive study of the performance and behavior of Large Multimodal Models (LMM)s under white-box visual adversarial attack.

Takeaways:

- LMMs are generally vulnerable to visual adversarial perturbations.
- Visual adversarial attacks are not *universal:* they are less effective when the query targets different visual contents.
- Adding additional textual context improves LMMs' robustness against visual adversarial input

Setup



Empirical Results

While the model completely misinterprets the object, it can still correctly answer questions that are not directly related with what's being attacked

Query: What is this image about?

LLaVA: Two sheep are standing on a ledge, looking over a wall.. ✓

LLaVA(adv): A group of people are sitting in a tub, with one person holding a toothbrush. ✗

Query: What animal is in the image?

LLaVA: Sheep. ✓

LLaVA(adv): None. ✗

Query: Is the photo taken indoor or outdoor?

LLaVA(adv): Outdoor. ✓

Query: Is there a tree in the image?

LLaVA(adv): Yes. ✓

Query: What is in the background of the image?

LLaVA(adv): Trees. ✓

Query: Sheep have a head with two large, curved horns, with a woolly coat that can vary in color..
Question: What animal is in the image?

LLaVA(adv): Sheep. ✓

Query: What is this image about?

LLaVA: a group of cows laying down in a barn. ✓

LLaVA(adv): A row of colorful wooden benches with Asian writing on them. ✗

Q1: Why is there a gap between the roof and wall?

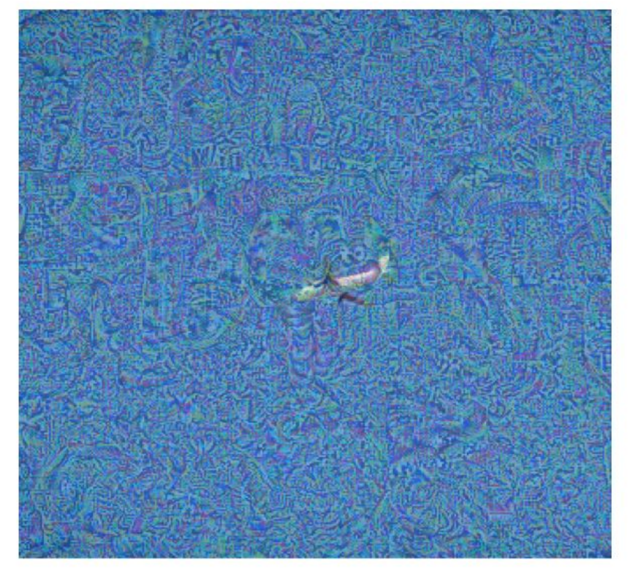
LLaVA(adv): Ventilation. ✓

Q2: Why is the cow laying down?

LLaVA(adv): Resting. ✓

Q3: Is it daylight in this picture?

LLaVA(adv): Trees. ✓



Query: What is this image about?

LLaVA: A large white airplane is flying through a clear blue sky. ✓

LLaVA(adv): A colorful, psychedelic patterned background features a pair of Elmo figurines from Sesame Street ✗

Q1: Was this taken at sunset?

LLaVA(adv): No. ✓

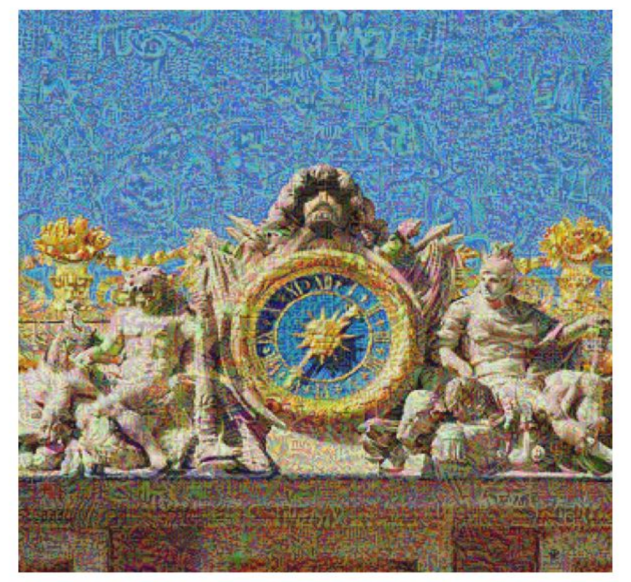
Q2: Is there a seagull?

LLaVA(adv): No. ✓

Q3: What color is the plane?

LLaVA: White. ✓

LLaVA(adv): Blue. ✗



Query: What is this image about?

LLaVA: The image features a large clock on the side of a building, surrounded by statues of men and animals.✓

LLaVA(adv): A man is sitting on a bench with two other people, one of whom is a cat. ✗

Q1: How many statues can you see?

LLaVA(adv): 5. ✓

Q2: Is it night time?

LLaVA(adv): No. ✓

Q3: What color is the clock face?

LLaVA(adv): Blue. ✓

VQA

Table 1. Accuracy percent dropped post-attack comparing to pre-attack. We can observe LMMs' VQA accuracy drops much less than that of their visual encoders'.

	LLaVA1.5		BLIP2-T5		InstructBLIP	
dataset	VQA	Vis.Enc.	VQA	Vis.Enc.	VQA	Vis.Enc.
MME	29.3	97.3	32.7	97.0	29.7	97.0
POPE	19.0	99.3	19.3	98.3	25.7	98.3
ScienceQA	4.4	98.7	5.8	99.7	5.1	99.7
TextVQA	23.3	99.3	25.3	99.7	39.3	99.7
VQAV2	27.4	97.0	29.3	99.0	35.0	99.0

VQA accuracy drops the most when questions are asking "what".

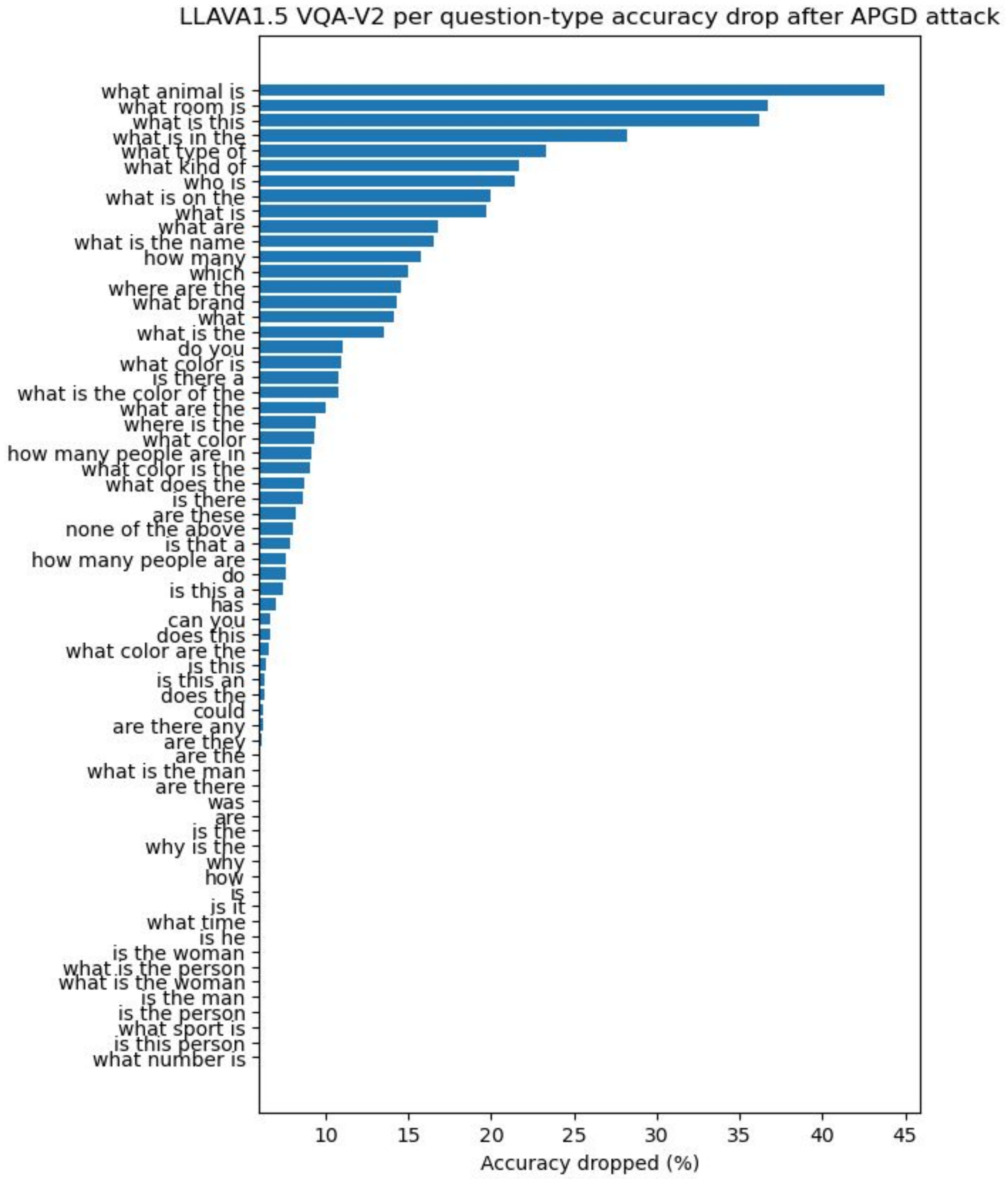


Image Captioning/Class.

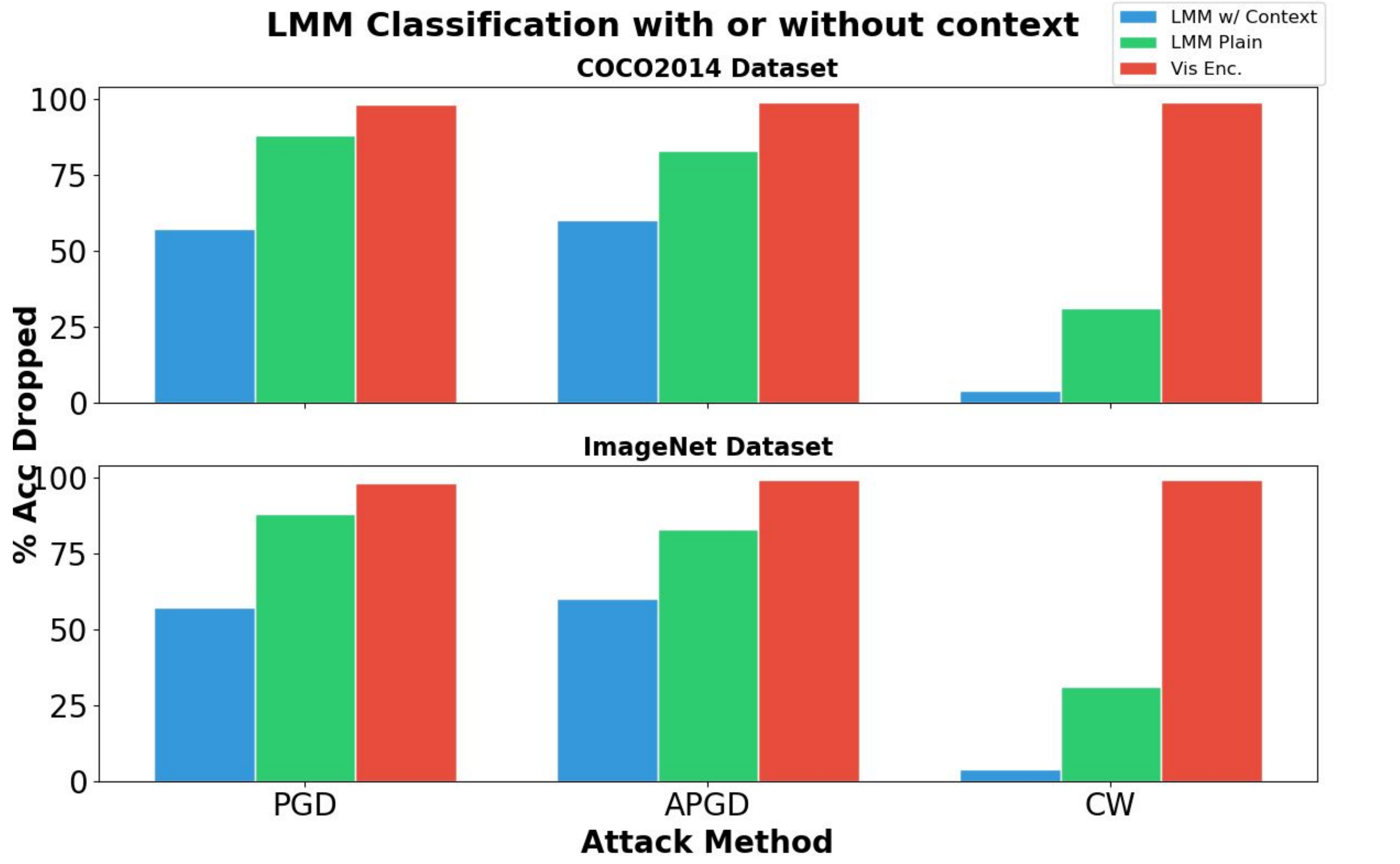
Table 2. Accuracy percent dropped post-attack comparing to pre-attack. Comparing to VQA, LMM accuracy drops are much more obvious on Image classification and caption retrieval.

Dataset	LLaVA	BLIP2	InstructBLIP	CLIP	BLIP
	Caption Retrieval Acc. Drop (%)				
COCO2014	62	99	78	98	78
	Image Classification Acc. Drop (%)				
COCO2014	31	75	86	98	100
Food101	79	83	75	96	100
StanfordCars	75	99	70	100	99

Context aids LMM Robustness against adversaries

Adding context of object description helps improve LMMs' robustness against adversarial visual inputs.

We can thus 'decompose' the query into a series of existential questions, each with the corresponding object context descriptio, and select the answer based on highest token probability.


LMM Classification with or without context
LMM w/ Context
LMM Plain
Vis Enc.
COCO2014 Dataset
ImageNet Dataset
% Acc Dropped
PGD
APGD
CW
Attack Method