On the Robustness of Large Multimodal Models Against Image Adversarial Attacks

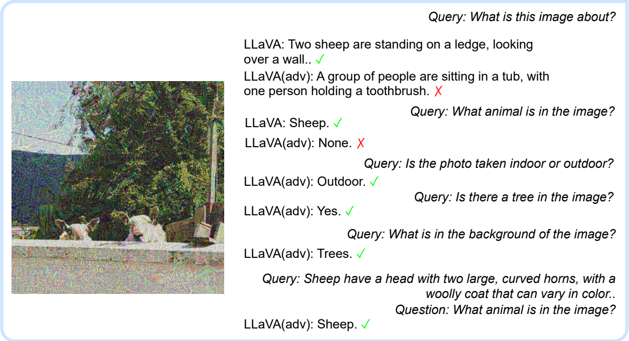
Xuanming Cui[†] xu979022@ucf.edu
Alejandro Aparcedo[†] aaparcedo@ucf.edu
Young Kyun Jang kyun0914@gmail.com
Ser-Nam Lim[†] sernam@ucf.edu

Abstract

Recent advances in instruction tuning have led to the development of State-of-the-Art Large Multimodal Models (LMMs). Given the novelty of these models, the impact of visual adversarial attacks on LMMs has not been thoroughly examined. We conduct a comprehensive study of the robustness of various LMMs against different adversarial attacks, evaluated across tasks including image classification, image captioning, and Visual Question Answer (VQA). We find that in general LMMs are not robust to visual adversarial inputs. However, our findings suggest that context provided to the model via prompts—such as questions in a QA pair—helps to mitigate the effects of visual adversarial inputs. Notably, the LMMs evaluated demonstrated remarkable resilience to such attacks on the ScienceQA task with only an 8.10% drop in performance compared to their visual counterparts which dropped 99.73%. We also propose a new approach to real-world image classification which we term query decomposition. By incorporating existence queries into our input prompt we observe diminished attack effectiveness and improvements in image classification accuracy. This research highlights a previously under explored facet of LMM robustness and sets the stage for future work aimed at strengthening the resilience of multimodal systems in adversarial environments.

1. Introduction

Large Multi-modal Models (LMMs) have demonstrated remarkable abilities in a range of applications, from image classification and Visual Question Answering (VQA) to image captioning and semantic segmentation [1, 13, 22, 23, 28]. These models excel in generalizing to new domains with data-efficient solution, a feat attributed to advancements in Instruction Tuning [42]. Such techniques, traditionally applied to text-only models, have now been extended to multi-modal models, opening new avenues for efficient fine-tuning with significantly less data [13, 28].



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. ✓

Figure 1. QA pairs for LLaVA [28] given an adversarial image. "LLaVA" and "LLaVA(adv)" refer to LLaVA's response to the user query with clean and adversarial image, respectively. For the readers, there are two sheep in the scene, and the adversarial attack was based on maximizing the distance between the image and the text "a photo of a *sheep*". In the first two QA pairs, we can see that LLaVA(adv)'s answer is completely wrong. However, it can still answer the following questions correctly, because they are not pertinent to the object being attacked (sheep). Also note the contrast between the second and last QA pairs. LLaVA(adv) answers the question correctly after additional context has been provided. These observations help drove some of the findings in this paper. Source: COCO 2014 [26]

Despite the recent advancements in LMMs, the impact of adversarial examples still remains under explored. Typically adversarial examples are generated end-to-end, targeting the final loss of the whole model, and focusing on a single modality. However, in the era of combining different pre-trained models with additional projectors or adaptors [8, 28, 44], it is imperative to reevaluate the effectiveness of these adversarial approaches. For example, let's consider LLaVA [28] which uses CLIP as its visual component and LLAMA as text component (with some additional projector to bridge the gap), will an attack on one of the two components compromise its overall performance?

From a practical perspective, given the substantial size of LMMs, attacking the entire model is often prohibitively expensive [7], making the above question an increasingly important one to answer since traditional adversarial attacks

[†] University of Central Florida

are better developed and computationally cheaper. Specifically, in this paper, we question the efficacy of adversarial attacks against visual encoders when they serve as input to subsequent LLMs. This gap in understanding raises critical questions about the susceptibility of LMMs to adversarial attacks, especially when only the visual encoder is targeted. Given their sophisticated dual-model composition, the question arises: *Can an attack on the visual encoder effectively compromise the entire LMM?*

Recent works [7, 35] on visual adversarial attacks against LMMs typically focus on the safety and alignment aspects of the model. For example, Qi et al [35] and Carlini et al [7] both show that it is possible to generate a visual adversarial example that "jailbreak" the LMM. Nonetheless, a systematic study on the impact of visual adversaries on LMMs is still missing.

We conduct a comprehensive analysis on the robustness of current LMMs under various adversarial attacks, tasks and datasets. Our investigation reveals that LMMs are not robust to adversarial visual perturbations in contexts where no additional textual information is provided, such as in COCO[26] classification (without context) or COCO captioning tasks. Conversely, the presence of context seems to bolster LMM robustness, as seen in tasks like COCO classification (with context). In cases where the attack does not directly target the core aspects of the task, such as in VQA, LMMs display a degree of inherent robustness. This paper reveals the following findings:

- LMMs are generally vulnerable to adversarial visual perturbations, even if such perturbations are generated only w.r.t. the visual model.
- Compared to classification and caption, LMMs demonstrate better robustness in VQA tasks. Particularly, we find that visual attacks are less effective when the VQA question query involves different visual contents from what is being attacked.
- Adding additional textual context notably improves LMMs' robustness against visual adversarial input.
- Based on the above findings, we devise a context-augmented image classification scheme that shows non-trivial increase in robustness.

2. Related Work

Large Multimodal Models (LMMs). Large Multimodal Models (LMMs)[4, 8, 23, 28, 44] typically comprise a visual model, a pre-trained Large Language Model (LLM), and a projector model designed to bridge the modality gap between images and text. Prominent among these models are LLaVA[28] and InstructBLIP [13], which represent the current state-of-the-art in LMMs. LLaVA integrates the CLIP visual encoder with the Vicuna LLM [10], employing a simple linear projector subsequent to the visual model for transforming visual representations into the language embedding space. Conversely, BLIP2-based models [13, 23, 44] utilize the EVA-CLIP visual encoder, alongside a Q-former equipped with learnable query vectors to bridge the visual and textual modalities. Both LLaVA and BLIP2-based models, among others, have demonstrated remarkable capabilities in a variety of vision-language tasks, underscoring their versatility and effectiveness.

Adversarial attacks. Adversarial attacks are designed to subtly manipulate inputs in a way that is typically imperceptible to humans, yet can lead neural networks to produce erroneous outputs [3, 5, 6, 12, 32, 39]. These attacks are broadly classified into two categories: white-box attacks [3, 6, 18, 39], where the adversary has complete access to the model parameters, and black-box attacks [34, 38], where the adversary possesses limited information such as output logits or labels. In particular, transfer-based attacks leverage gradients from a surrogate model under white-box condition, which are likely transferable to the target black-box model [15, 29, 33, 34]. Such transferability thus remain as an critical model vulnerability.

While the primary focus of adversarial attack research has historically been on image classification, recent studies have demonstrated the feasibility of constructing adversarial examples in textual domains. These examples can be generated either heuristically [2, 21, 24] or through discrete optimization techniques [16, 40].

LMMs and Adversarial Examples. While extensive research has been conducted on adversarial attacks in both visual and textual domains, the impact of these attacks on current LMMs remains relatively unexplored. Recent studies [7, 30, 35, 36, 41, 45] demonstrate the feasibility of creating adversarial examples that effectively "jailbreak" LMMs from both visual [7, 35] and textual [30, 36, 41, 45] inputs, using either gradient-based approaches [7, 35] or prompt engineering [30, 36, 41]. These examples are capable of inducing LMMs to produce harmful content, thereby bypassing the safety measures implemented during model alignment, such as instruction tuning or Reinforcement Learning from Human Feedback (RLHF). However, while these studies predominantly address the safety concerns, potential harmfulness, and the associated dangers of LMMs, our research shifts the focus towards systematically examining the accuracy of LMMs in performing various tasks under the influence of visual adversarial attacks.

3. Method

3.1. Threat Model

In this study, we focus on gradient-based white-box adversarial attacks [6, 12, 32]. These methods hinge on the computation of the gradient to ascertain the most effective direction in which to modify the input so as to deceive the model, while satisfying the L_p constraint. Formally, given input-label pair x, y and the model denoted by f, we want to find the adversarial perturbation δ s.t. $f(x+\delta) \neq y$ confined to some L_p bounds. For PGD, we maximize $L(f(x+\delta), y)$ while satisfying $||\delta||_\infty < \epsilon$, where ϵ is the radius of the L_∞ ball, and L is the Cross-Entropy loss in our case. For CW, we maximize $||\delta||_p + c \cdot g(x+\delta)$, subject to $x+\delta \in [0,1]$, where $g(x+\delta) = \max(f(x+\delta)_y - \max\{f(x+\delta)_i : i \neq y\}, -\kappa)$, and κ is the confidence parameter.

3.2. Attacks

We choose PGD and CW as two representatives of strong gradient-based attacks, along with APGD as a variant of PGD. Additionally, we experiment with two parameter settings of each attack: normal and strong, based on perceptibility of the perturbations. Under the normal setting, we set the constraint for CW to 20, and epsilon for PGD/APGD to 8/255, as used in prior works [3, 43]. Under the strong setting, we set epsilon for PGD and APGD to 0.2, and constraint to 100 for CW. All the attacks are generated solely w.r.t. the image encoder, leaving the LLM untouched. Detailed parameters can be found in Table 1.

Figure 2 shows a sample adversary generated using different attack methods and under different degree of attack strength. In the normal setting, the adversarial perturbation is almost imperceptible, but become obvious under strong setting for PGD and APGD. Perturbations generated by CW remains imperceptible even under the strong setting. For brevity, in the follow sections, we use N and S to represent normal and strong setting, respectively. For example, APGD-S stands for APGD attack under strong setting.

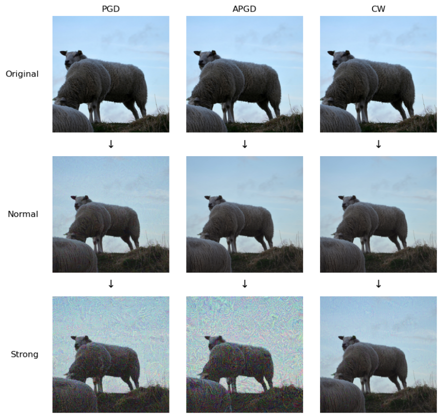


Figure 2. A sample CLIP's adversarial image, generated by PGD, APGD and CW, under Normal and Strong attack parameter settings. Image source: COCO 2014val. Note that under strong attack, the adversarial perturbations become very obvious under PGD and APGD, and are expected to cause a higher degree of performance degradation.

Method	steps	step size	ϵ	Dist.	c	κ
			Normal			
PGD	20	2/255	8/255	L_∞	-	-
APGD	20	-	8/255	L_∞	-	-
CW	50	0.01	-	L_2	20	0
			Strong			
PGD	40	2/255	0.2	L_∞	-	-
APGD	40	-	0.2	L_∞	-	-
CW	75	0.05	-	L_2	100	0

Table 1. Parameters for the attacks under Normal (N) and Strong (S) settings. Dist. refers to distance measure, c and κ refers to the constraint and confidence parameter in [6].

3.3. Models

In our study, we selected three state-of-the-art LMM models for evaluation: LLaVA1.5[27] integrated with the Vicuna13B language model, BLIP2 combined with the Flan T5 XXL[11] language model, and InstructBLIP [13], also utilizing Vicuna13b. These models exhibit distinct characteristics in their configurations. LLaVA1.5 and InstructBLIP both employ the Vicuna13B language model; however, they differ in their image encoders and methodologies for merging image and text encodings, with LLaVA1.5 directly inserting the projected visual tokens into text tokens, versus InstructBLIP's Q-former architecture. BLIP2 and InstructBLIP share similar image encoders and the Q-former architecture but diverge in their language model choices and training protocols: BLIP2 employs the Encoder-Decoder based Flan T5 XXL, while InstructBLIP uses the Decoder-only Vicuna13B. We believe that such a selection of models allows for a diverse yet controlled set of experimental conditions. In the rest of the paper, we use LLaVA, BLIP2-T5 and InstructBLIP to refer to LLaVA1.5 Vicuna13B, BLIP2 Flan T5XXL, and InstructBLIP Vicuna13B, respectively.

3.4. Tasks & Adversarial generation

We consider three popular visual tasks for evaluating visual adversarial impact on LMMs: image classification, caption retrieval and VQA. Since we are interested in LMMs' robustness against visual adversaries, we generate adversarial samples w.r.t. the image encoder of the LMM: CLIP image encoder for LLaVA, EVA-CLIP image encoder for BLIP2 and InstructBLIP. We use CLIP text encoder and the text encoder from BLIP's Q-former to compute the text embeddings for their corresponding image encoder. For PGD and APGD, we maximize the Cross-Entropy loss between the model logits and the ground-truth label. For CW, we minimize the sum of the l_2 distance of the perturbation δ and the f-function from the original paper [6]. Detailed procedures for task-specific adversarial generation are given below.

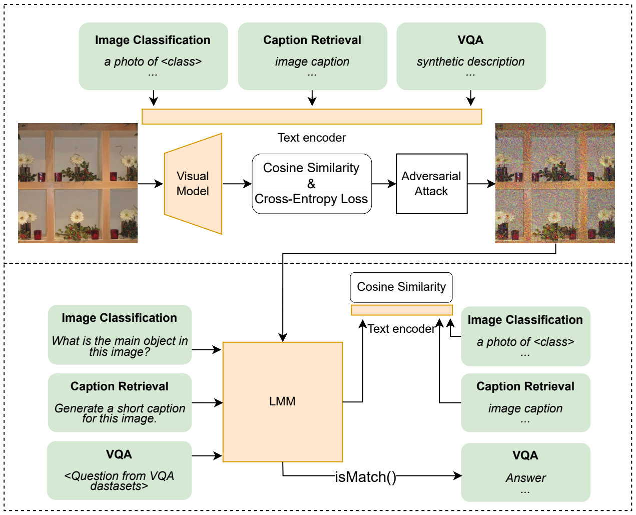


Figure 3. Overview of our procedure for attack generation and evaluation over image classification, caption retrieval, and VGA. Top: overview of attack generation for the three tasks; bottom: evaluation procedure for LMM on the three tasks.

3.4.1 Image Classification

We use COCO [26] 2014 validation split (2014val), with class annotations from [20], to evaluate robustness on classification. We first use the text encoder to encode the text class labels in the format of "a photo of <class>". Then, we compute the class-wise cosine-similarity between the image encodings and encoded class labels and use the result as the class logits for adversarial generation and evaluation. To evaluate LMMs on classification, we first prompt LMMs to generate a one-word response of the main object in the image. For LLaVA we use the prompt: "What is the main object in this image?\nAnswer in a single word or phrase." For BLIP2-T5 and InstructBLIP we use "Question: what is the main object in this image? Short answer: ". We then format the answer with "a photo of <answer>", encode with the text encoder and compute cosine similarity against the class label encodings to perform classification.

3.4.2 Caption retrieval

We use COCO captioning dataset [9] 2014val for evaluating caption retrieval robustness. To generate visual adversarial samples for caption retrieval, we first use the text encoder to encode 5 captions per image, and then use their mean as the text encodings for each image. Then, we compute cosine similarity between image and text encodings and use the result as the image-wise logits for adversarial generation. To evaluate the caption retrieval for CLIP and EVA-CLIP encoders, we compute cosine similarity between the image encodings and all captions' text encodings to perform retrieval. To evaluate caption retrieval for LMMs, we first prompt LMMs to generate a caption for the image. For LLaVA we use the prompt: "Describe this image in a short sentence." For BLIP2-T5 and InstructBLIP we use the prompt: "Question: what is this image about? Short answer: " Then we encode the generated caption and compute cosine similarity against all captions' text encodings to perform retrieval.

3.4.3 VQA

We evaluate LMM robustness on five popular VQA datasets: VQA V2 [19], ScienceQA-Image [31], TextVQA [37], POPE [25] and MME [17]. For VQA V2, we follow the same adversarial generation procedure as in classification task. For all other datasets, since no ground-truth label is present, we first prompt LLaVA with "What is this image about?\nAnswer in one sentence." to generate synthetic caption for each image, and follow the same procedure in caption retrieval task for generating adversaries. We follow the official evaluation procedures for each VQA datasets.

4. Experimental Results and Analysis

We show our experimental results and analysis in the following sections. We report both LMMs' accuracy as well as the image encoder's accuracy on the task that was used to generate adversaries. We adopt the notations Pre, Post_N and Post_S to refer to accuracy for pre-attack, post-attack under normal setting, and post-attack under strong setting, respectively.

4.1. Are LMMs Robust Against Adversarial Visual Input?

To investigate the impact of adversarial visual inputs on LMMs, our initial analysis focuses on the caption retrieval task. This task serves as a measure of the LMMs' overall comprehension of visual inputs. The results of this analysis, conducted on COCO 2014val, are presented in Table 2 – please refer to the caption to comprehend the significance of each metric. Under the third section, the data distinctly illustrates a significant decrease in post-attack accuracy across all three LMMs when subjected to both PGD and APGD attacks, under both normal and strong settings. For instance, under PGD-N, the Top-1 recall rate for InstructBLIP declines to 1.9%, and further diminishes to 0.18% under PGD-S. Since these accuracies are roughly on the same level as the CLIP/EVA-CLIP accuracy post attack, as shown under the second section of the table, the additional LLM appended to the image encoder did not bring notable robustness, indicating that LMMs lack robustness against visual adversarial perturbations. In other words, the visual perturbations are capable of substantially undermining the LMMs' effectiveness, even though they are not gen-

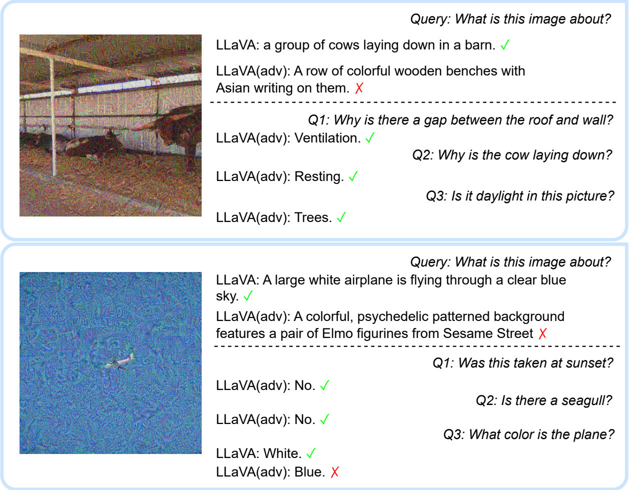
Model	Attack	Pre	Post_N	Post_S
		Visual Encoder Acc @1 (%)		
CLIP	PGD	63.32	11.78(-81)	0.48(-99)
CLIP	APGD	63.32	4.2(-93)	0.02(-100)
CLIP	CW	63.32	13.86 (-78)	0.94(-99)
EVA-CLIP	PGD	73.18	1.02(-99)	0.0(-100)
EVA-CLIP	APGD	73.18	0.46(-99)	0.0(-100)
EVA-CLIP	CW	73.18	19.43 (-73)	3.74(-95)
		Image-to-Text Recall @1 (%)		
CLIP	PGD	57.72	10.4(-82)	0.4(-99)
CLIP	APGD	57.72	12.92(-78)	7.44(-87)
CLIP	CW	57.72	34.94(-39)	24.94(-57)
EVA-CLIP	PGD	64.06	1.06(-98)	0.06(-100)
EVA-CLIP	APGD	64.06	9.14(-86)	8.32(-87)
EVA-CLIP	CW	64.06	42.06(-34)	31.72(-50)
		LLM Answer-to-Text Recall @1 (%)		
LLaVA	PGD	36.58	13.1(-64)	3.76(-90)
LLaVA	APGD	36.58	15.7(-57)	7.88(-78)
LLaVA	CW	36.58	32.96(-10)	29.84(-18)
BLIP2-T5	PGD	32.34	1.4(-96)	0.1(-100)
BLIP2-T5	APGD	32.34	4.52(-86)	3.62(-89)
BLIP2-T5	CW	32.34	23.12(-29)	17.02(-47)
InstructBLIP	PGD	37.82	1.9(-95)	0.18(-100)
InstructBLIP	APGD	37.82	5.56(-85)	4.3(-89)
InstructBLIP	CW	37.82	27.44(-27)	20.74(-45)

Table 2. Top-1 caption retrieval result for COCO caption 2014 validation dataset. Refer to Sec. 3.4.2. "Visual Encoder Accuracy" refers to CLIP/EVA-CLIP accuracy on successfully retrieving captions that are closed to the mean caption encoding given the image encoding. "Image-to-Text Recall @1" is recall@1 of retrieving correctly one of the five captions for the given image. LLM Answer-to-Text recall is the same except the query is the LMMs' answers. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

erated through an end-to-end process on the LMMs' text generation loss.

4.2. Evaluating LMMs' VQA Performance

In this section, we detail the experimental outcomes of the LMMs in VQA tasks under adversarial visual attacks. The primary results are summarized in Table 3. Our results indicate a noteworthy deviation from what we have observed about the caption retrieval task in Sec. 4.1, which did not show that LMMs possess any robustness against visual adversaries. Based on the results from Table 3, all three LMMs being evaluated exhibit considerable resilience in various VQA datasets, despite the significant decrease in adversarial accuracy of their corresponding visual encoders, as shown under the "Visual Encoder Accuracy" columns. For instance, with the ScienceQA dataset, the Post_N "Visual Encoder Accuracy" plummeted below 1% for all three types of attacks, and for both the CLIP and BLIP visual encoders. However, the accuracy of all three LMMs decreased by less than 7% compared to their pre-attack accuracy.



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. ✗

Figure 4. Two sample adversarial images from COCO 2014val, generated under APGD Post_S. "LLaVA" and "LLaVA(adv)" refer to LLaVA's responses using the clean Pre-attack and post-attack image, respectively. Above the dotted line in each cell, we query LLaVA for the general description; below the dotted line are questions taken from VQA V2 dataset.

What could be the cause of such discrepancies in LMMs' robustness between the VQA and caption retrieval tasks? We make two conjectures:

1. The robustness of LMMs depends on whether the query is about what is being attacked. Since the attack target for generating visual adversarial samples is what is being described in the image description, then intuitively those aspects not mentioned in the description shall be less affected by the attack.
2. Additional contexts (e.g., contexts in ScienceQA's questions) aid in LMMs' robustness.

We will experimentally support the two claims in the following sections.

4.3. Visual Adversarial Attacks are not Universal to LMMs

In this section, we present an empirical analysis demonstrating that while LMMs are not inherently resilient to visual adversarial attacks, as evidenced by their performance in caption retrieval tasks, they are capable of delivering correct responses when the query's focus differs from the target of the attack. To illustrate this, we take the Visual Question Answering (VQA) V2 dataset as a case study. Here, we generate adversarial images using the text label "a photo of <class>", with the attack primarily aimed at the central object of the image. We observe that the adversarial attack's effectiveness is heightened when the query, during evaluation, pertains to the same target – the principal object in the

				VQA Acc (%)			Visual Encoder Acc (%)	
Model	Dataset	Attack	Pre	Post$_\text{N}$	Post$_\text{S}$	Pre	Post$_\text{N}$	Post$_\text{S}$
LLaVA	ScienceQA(image)	PGD	71.59	68.77 (-3)	64.75 (-9)	42.92	10.08 (-77)	0.92 (-98)
LLaVA	ScienceQA(image)	APGD	71.59	69.81 (-2)	68.22 (-4)	42.83	5.68 (-87)	0.06 (-99)
LLaVA	ScienceQA(image)	CW	71.59	71.69 (+0.1)	71.34 (-0.3)	42.95	12.76 (-70)	0.03 (-99)
BLIP2-T5	ScienceQA(image)	PGD	74.71	69.71 (-6)	63.06 (-15)	46.40	1.05 (-98)	0.00 (-100)
BLIP2-T5	ScienceQA(image)	APGD	74.71	73.62 (-1)	72.88 (-2)	46.40	1.02 (-98)	0.00 (-100)
BLIP2-T5	ScienceQA(image)	CW	74.71	74.71 (-0)	74.37 (-0.4)	46.40	12.92 (-72)	0.03 (-99)
InstructBLIP	ScienceQA(image)	PGD	45.08	40.66 (-10)	39.67 (-12)	46.40	1.05 (-98)	0.00 (-100)
InstructBLIP	ScienceQA(image)	APGD	45.08	42.75 (-5)	42.34 (-3)	46.40	1.02 (-98)	0.00 (-100)
InstructBLIP	ScienceQA(image)	CW	45.08	44.88 (-0.4)	43.93 (-0.3)	46.40	12.92 (-72)	0.03 (-99)
LLaVA	VQA V2	PGD	78.43	64.38 (-18)	51.22 (-35)	89.21	31.00 (-65)	6.56 (-93)
LLaVA	VQA V2	APGD	78.43	67.41 (-14)	44.60 (-43)	89.21	30.01 (-66)	0.15 (-99)
LLaVA	VQA V2	CW	78.43	76.79 (-2.1)	75.16 (-4.1)	89.21	44.92 (-50)	0.04 (-99)
BLIP2-T5	VQA V2	PGD	66.94	50.60 (-24)	42.43 (-37)	94.13	19.12 (-80)	0.23 (-99)
BLIP2-T5	VQA V2	APGD	66.94	52.51 (-22)	40.69 (-39)	94.13	14.71 (-84)	0.01 (-99)
BLIP2-T5	VQA V2	CW	66.94	63.69 (-4.8)	58.75 (-12)	94.12	51.60 (-45)	0.06 (-99)
InstructBLIP	VQA V2	PGD	76.07	56.33 (-26)	42.77 (-44)	94.13	19.12 (-80)	0.23 (-99)
InstructBLIP	VQA V2	APGD	76.07	58.83 (-22)	39.60 (-48)	94.13	14.71 (-84)	0.01 (-99)
InstructBLIP	VQA V2	CW	76.07	73.02 (-4.0)	66.10 (-13)	94.12	51.60 (-45)	0.06 (-99)
LLaVA	TextVQA	PGD	62.14	51.44 (-17)	40.27 (-35)	69.32	10.30 (-85)	0.41 (-99)
LLaVA	TextVQA	APGD	62.14	54.23 (-12)	42.88 (-31)	69.32	6.73 (-90)	0.00 (-100)
LLaVA	TextVQA	CW	62.14	60.88 (-2)	59.71 (-4)	69.38	18.00 (-74)	0.03 (-99)
BLIP2-T5	TextVQA	PGD	45.14	38.46 (-14)	29.82 (-34)	68.97	0.44 (-99)	0.00 (-100)
BLIP2-T5	TextVQA	APGD	45.14	39.94 (-11)	32.41 (-28)	68.97	0.63 (-99)	0.00 (-100)
BLIP2-T5	TextVQA	CW	45.14	44.28 (-2)	38.58 (-14)	69.01	12.51 (-82)	0.03 (-99)
InstructBLIP	TextVQA	PGD	35.23	26.99 (-23)	18.11 (-48)	68.97	0.44 (-99)	0.00 (-100)
InstructBLIP	TextVQA	APGD	35.23	28.00 (-20)	20.10 (-43)	68.97	0.63 (-99)	0.00 (-100)
InstructBLIP	TextVQA	CW	35.23	33.95 (-3)	25.46 (-27)	69.01	12.51 (-82)	0.03 (-99)
LLaVA	POPE	PGD	85.55	73.13 (-14)	58.97 (-31)	80.00	7.20 (-91)	0.2 (-99)
LLaVA	POPE	APGD	85.55	73.80 (-13)	65.00 (-24)	80.00	4.40 (-94)	0.00 (-100)
LLaVA	POPE	CW	85.55	83.07 (-2)	83.27 (-2)	80.00	21.40 (-73)	0.80 (-99)
BLIP2-T5	POPE	PGD	77.10	62.67 (-18)	55.50 (-28)	87.40	0.00 (-100)	0.00 (-100)
BLIP2-T5	POPE	APGD	77.10	65.20 (-15)	55.30 (-28)	87.40	0.20 (-99)	0.00 (-100)
BLIP2-T5	POPE	CW	77.10	75.87 (-1)	75.20 (-2)	87.40	15.80 (-81)	3.80 (-95)
InstructBLIP	POPE	PGD	82.83	64.57 (-22)	52.20 (-37)	87.40	0.00 (-100)	0.00 (-100)
InstructBLIP	POPE	APGD	82.83	66.93 (-19)	52.00 (-37)	87.40	0.20 (-99)	0.00 (-100)
InstructBLIP	POPE	CW	82.83	80.93 (-2)	80.27 (-3)	87.40	15.80 (-82)	3.80 (-95)
LLaVA	MME	PGD	1,536	1,187 (-22)	927 (-39)	65.79	12.25 (-81)	0.91 (-98)
LLaVA	MME	APGD	1,536	1,283 (-16)	818 (-46)	65.79	7.29 (-88)	0.10 (-99)
LLaVA	MME	CW	1,536	1,521 (-1)	1491 (-3)	65.79	17.51 (-73)	3.04 (-95)
BLIP2-T5	MME	PGD	1,114	759 (-32)	591 (-47)	73.38	2.43 (-96)	0.00 (-100)
BLIP2-T5	MME	APGD	1,114	777 (-30)	628 (-43)	73.38	2.02 (-97)	0.00 (-100)
BLIP2-T5	MME	CW	1,114	1058 (-5)	1026 (-8)	73.38	18.62 (-74)	6.17 (-91)
InstructBLIP	MME	PGD	1,248	704 (-30)	703 (-43)	73.38	2.43 (-96)	0.00 (-100)
InstructBLIP	MME	APGD	1,248	1,002 (-19)	751 (-40)	73.38	2.02 (-97)	0.00 (-100)
InstructBLIP	MME	CW	1,248	1,205 (-3)	1,170 (-6)	73.38	18.62 (-74)	6.17 (-91)

Table 3. Results on VQA datasets. We attack CLIP and EVA-CLIP visual encoders to generate adversarial examples for LLaVA and BLIP2-T5/InstructBLIP, respectively. Adversarial examples are used as input image along with question as input text. "VQA Accuracy" refers to the performance of each LMM; "Visual Encoder Accuracy" refers to the accuracy of the visual encoder on image-to-text retrieval, which is used for generating visual adversaries for VQA. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

image. Conversely, the attack's impact diminishes when the query relates to different aspects of the image.

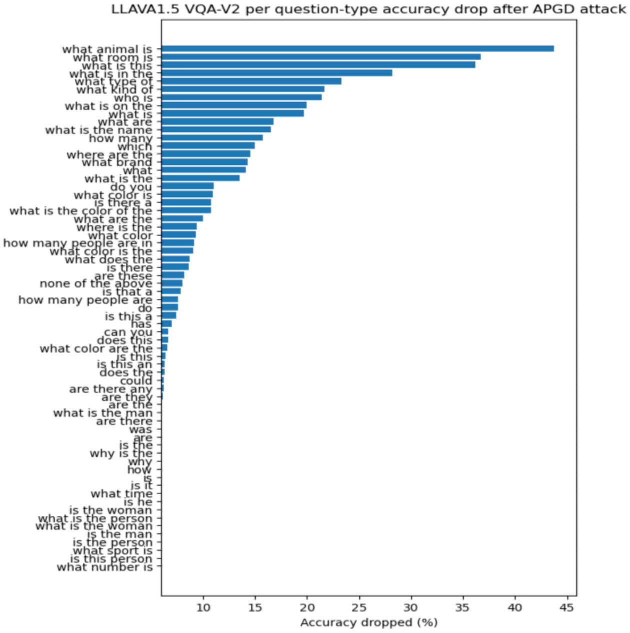


Figure 5. LLaVA's VQA accuracy drop after APGD attack under the normal attack setting. Y-axis represents question types, and X-axis represents accuracy dropped (%).

In Figure 4, we show LLaVA's responses to queries on two adversarial images under APGD-S. When querying about the general description of the image, it is clear that LLaVA's post-attack answers are completely deviated from what the image is about; however, below the dotted line, LLaVA can still answer most questions correctly. We conjecture such phenomenon is either because LLaVA can "guess" the answer directly from the context (e.g., Q2-top "Why is the cow laying down?" – "Resting"). This is coherent with LMMs' high "robustness" on the ScienceQA dataset, in which the texts themselves are often sufficient to find the answer. On the other hand, it is because these questions are not directly querying the object or its attributes, but rather the peripheral aspects of the image (e.g., Q1-bottom "Was this taken at sunset?"). The only incorrectly-answered question is Q3-bottom "What color is the plane?". LLaVA answers it incorrectly as the query is asking about the object attribute (color), which has been corrupted by the attack.

In Figure 5, we plot LLaVA's per-question type accuracy drop under APGD-N. We can clearly see that accuracy drops the most on questions asking 'What' – what room/animal/color – about the object and its direct attributes. The accuracy drop quickly diminishes for question asking 'Is/Has/Can' etc. These questions are typically querying the peripheral aspects of the image instead of the main object, and actually require more complex reasoning and understanding to answer correctly, yet they boast much lower accuracy drop. This result reaffirms our conjecture that LMMs are robust when the question is not querying what is being attacked.

Model	Attack	Pre@1	Post$_N$@1	Post$_S$@1
		Visual Encoder Acc (%)		
CLIP	PGD	89.21	31.0(-65)	6.56(-93)
CLIP	APGD	89.21	30.01(-66)	0.15(-100)
CLIP	CW	89.21	44.92(-50)	0.04(-100)
EVA-CLIP	PGD	94.13	19.12(-80)	0.23(-100)
EVA-CLIP	APGD	94.13	14.71(-84)	0.01(-100)
EVA-CLIP	CW	94.12	51.6(-45)	0.06(-100)
		LMM Acc (%)		
LLaVA	PGD	87.51	48.25(-45)	22.58(-74)
LLaVA	APGD	87.51	52.06(-41)	8.11(-91)
LLaVA	CW	87.51	80.64(-8)	77.1(-12)
BLIP2-T5	PGD	86.47	28.64(-67)	2.98(-97)
BLIP2-T5	APGD	86.47	31.39(-67)	2.37(-97)
BLIP2-T5	CW	86.47	70.11(-19)	58.85(-32)
InstructBLIP	PGD	89.89	21.09(-77)	3.66(-96)
InstructBLIP	APGD	89.89	22.35(-75)	2.18(-98)
InstructBLIP	CW	89.89	37.81(-58)	31.91(-64)
		LMM with Context Acc (%)		
LLaVA	PGD	93.74	73.62(-21)	57.06(-39)
LLaVA	APGD	93.74	72.61(-23)	37.65(-60)
LLaVA	CW	93.74	91.76(-2)	90.2(-4)
BLIP2-T5	PGD	97.67	87.54(-10)	94.92(-3)
BLIP2-T5	APGD	97.67	87.29(-11)	98.43(+1)
BLIP2-T5	CW	97.67	94.97(-3)	92.76(-5)
InstructBLIP	PGD	88.94	66.92(-25)	71.61(-19)
InstructBLIP	APGD	88.94	68.74(-23)	89.22(-0)
InstructBLIP	CW	88.94	84.92(-5)	82.51(-7)

Table 4. Top-1 image classification result on COCO 2014val. The first table section shows visual encoder accuracy, referring to CLIP/EVA-CLIP's accuracy on classification; second section shows LMMs' accuracy; third section show LMMs' accuracy, after the context is added to the query. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

4.4. Adding Context Improves LMM Robustness

To examine the effect of context on LMMs' robustness, we reuse the image classification task. We first ask LLaVA to generate a general one-sentence description for each class. We then insert the generated description corresponding to the correct object into the prompt for querying the LMMs about the main object in the image. Besides the additional context, everything else is kept the same.

Results are shown in Table 4. We observe that after adding a short sentence of context, the post-attack accuracy for all three LMM models increase by a large margin. In particular, the accuracy drop for BLIP2/InstructBLIP under PGD/APGD reduce to only 20%, as opposed to an average of 60% drop without context. Although the resulting accuracy is still not on par with the pre-attack accuracy, it still

		LMM Query Decomp. Acc(%)			LMM Plain Acc (%)			Visual Encoder Acc (%)		
Dataset	Attack	Pre	Post$_\text{N}$	Post$_\text{S}$	Pre	Post$_\text{N}$	Post$_\text{S}$	Pre	Post$_\text{N}$	Post$_\text{S}$
COCO	PGD	98.42	65.30 (-34)	35.88 (-64)	87.51	48.25 (-45)	22.58 (-74)	89.21	31.00 (-65)	6.56 (-93)
COCO	APGD	98.42	66.98 (-32)	23.88 (-76)	87.51	52.06 (-41)	8.11 (-91)	89.21	30.01(-66)	0.15 (-99)
COCO	CW	98.42	95.27 (-3)	93.68 (-5)	87.51	80.64 (-8)	77.10 (-12)	89.21	44.92 (-50)	0.04 (-99)
Imagenet	PGD	90.62	58.52 (-35)	29.96 (-67)	28.10	10.34 (-63)	3.44 (-88)	71.47	15.94 (-78)	1.07 (-98)
Imagenet	APGD	90.62	57.26 (-37)	27.16 (-70)	28.10	11.46 (-59)	4.72 (-83)	71.47	4.21 (-94)	0.01 (-99)
Imagenet	CW	90.62	87.72 (-3)	86.94 (-4)	28.10	21.00 (-25)	19.44 (-31)	71.47	11.62 (-84)	0.80 (-99)

Table 5. LLaVA classification accuracy on COCO 2014val and Imagenet 2012val. "LMM Query Decomp." refers to classification with context and query decomposition, as discussed in Sec. 4.5. "LMM plain" refers to classification without context and query decomposition. "Visual Encoder" refers to CLIP/EVA-CLIP's classification accuracy. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

suggests the efficacy of providing additional context against adversarial input, possibly by helping LLMs recover object attributes from the corrupted visual inputs and match with the correct object. This can be useful when the task is to identify the existence of some target objects (e.g., illicit object detection) from images that could be intentionally manipulated, and in the worst scenario, be adversarial.

Interestingly, we also observe that for BLIP2-T5 and InstructBLIP under PGD and APGD attacks, the Post$_\text{S}$ accuracy are higher than the normal setting. For APGD, they are even higher than the pre-attack accuracy. We conjecture this is due to the fact that APGD-S is too effective on EVA-CLIP (0.01% classification accuracy post-attack), and that BLIP2-T5 and InstructBLIP are solely relying on the object description to generate the answer while ignoring the adversarial visual input. The two LLMs therefore hallucinate the object description as the answer. However, although the post-attack accuracy is also low for CLIP (0.15% under APGD S.), we do not observe the same behavior for LLaVA. Possibly the reason is due to different ways LLaVA and BLIP combine the two modalities. While LLaVA takes visual input as standalone tokens, separately from text tokens, BLIP utilizes a Q-former, which blends two modalities together and therefore possibly outweighing the visual input signal with text's.

4.5. Towards Real-World Application: Context-Augmented Image Classification

In the previous section, we show that adding the correct object context enhances LMMs' robustness against adversarial images. In practice, the correct context is typically unknown. However, in the case of closed-world image classification, where the list of object classes are fixed, we can decompose each question into multiple existence questions. Each question queries the presence of one object class, along with the context corresponding to that object. Afterwards, we choose the object with the highest confidence from the LLM's final projection head. We term our approach query decomposition.

While this solution may appear to be brute-force, such a scheme is inherently able to support making each query in parallel, thereby improving the efficiency. Nevertheless, we would like to demonstrate how we can apply our findings towards real-world setting, and hope this approach presents a viable starting point that opens the door to future work. To see whether our proposed query decomposition may work, we conduct experiments using COCO 2014val and Imagenet [14] 2012 val. For each image, we randomly select 20 object classes while ensuring the correct object class is included. Results are shown in Table 5. We again observe noticeable improvements on robustness, just like in Table 4. For example, with the context and query decomposition under "LMM Query Decomp", the percent drops for COCO are mostly 10% smaller, and 20% smaller for Imagenet, comparing to post-attack accuracy drops without context as shown under "LMM plain Acc" columns.

Notably, when query decomposition is utilized to insert context, LMM's performance on ImageNet classification is greatly boosted. This can be seen by comparing the pre-attack performance under "LMM Query Decomp" and under "LMM Plain Acc".

5. Conclusion

In this study, we systematically evaluate the susceptibility of LMMs to visual adversarial inputs across a diverse array of tasks and datasets. Our findings suggests LMMs are highly vulnerable to visual adversarial attacks, even when such adversaries are crafted with respect to the visual model alone. On the other hand, we find that LMMs are "robust" when the query and attack target does not match. Such characteristics indicates the traditional task-specific adversarial generation techniques are not universally effective against current LMM, and points to the need for further research into new adversarial attack strategies, particularly in the context of zero-shot inference. Finally, we find adding context about the querying object improves LMMs' visual robustness. We therefore propose a strategy to decompose questions into multiple existence questions associated with the corresponding context, which achieved notable improvements in robustness on COCO and Imagenet classification.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[2] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, 2018. Association for Computational Linguistics. 2

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 2018. 2, 3

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[5] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. *Evasion Attacks against Machine Learning at Test Time*, page 387–402. Springer Berlin Heidelberg, 2013. 2

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 2, 3

[7] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2023. 1, 2

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. 1, 2

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 4

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020. 2

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[15] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 2

[16] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia, 2018. Association for Computational Linguistics. 2

[17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 4

[18] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 4

[20] Young Kyun Jang, Geonmo Gu, Byungsoo Ko, Isaac Kang, and Nam Ik Cho. Deep hash distillation for image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[21] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 2

[22] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation

via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 2

[24] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2022–2031, 2021. 2

[25] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 4

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 4

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2

[29] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. 2

[30] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2023. 2

[31] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 4

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2

[33] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016. 2

[34] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. page 506–519, New York, NY, USA, 2017. Association for Computing Machinery. 2

[35] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023. 2

[36] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks, 2023. 2

[37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 4

[38] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23:828–841, 2017. 2

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2

[40] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, 2019. Association for Computational Linguistics. 2

[41] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. 2

[42] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. 1

[43] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 3

[44] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1, 2

[45] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 2