On the Robustness of Large Multimodal Models Against Image Adversarial Attacks

Supplementary Material

Transfer	Attack	Vis.Enc. Acc (%) Post_N	Vis.Enc. Acc (%) Post_S	LMM Acc (%) Post_N	LMM Acc (%) Post_S
COCO					
$\text{BLIP}_v \rightarrow$ LLaVA	PGD	84.27 (-2)	64.97 (-27)	82.76 (-5)	66.77 (-24)
$\text{BLIP}_v \rightarrow$ LLaVA	APGD	85.09 (-2)	61.54 (-31)	82.89 (-5)	61.75 (-29)
$\text{BLIP}_v \rightarrow$ LLaVA	CW	86.57 (-2)	84.00 (-6)	84.56 (-3)	82.21 (-6)
CLIP $\rightarrow$ BLIP2	PGD	93.54 (-9)	90.28 (-4)	87.14 (-0)	83.79 (-3)
CLIP $\rightarrow$ BLIP2	APGD	93.29 (-9)	86.10 (-9)	87.03 (-1)	79.14 (-8)
CLIP $\rightarrow$ BLIP2	CW	93.64 (-9)	93.44 (-1)	87.44 (-0)	87.07 (-0)
CLIP $\rightarrow$ Ins.BLIP	PGD	93.54 (-5)	90.28 (-4)	89.28 (-1)	86.57 (-4)
CLIP $\rightarrow$ Ins.BLIP	APGD	93.29 (-5)	86.10 (-9)	89.13 (-1)	82.97 (-8)
CLIP $\rightarrow$ Ins.BLIP	CW	93.64 (-5)	93.44 (-1)	89.57 (-0)	89.21 (-1)
Food-101					
$\text{BLIP}_v \rightarrow$ LLaVA	PGD	66.96 (-26)	36.84 (-59)	17.24 (-46)	10.54 (-67)
$\text{BLIP}_v \rightarrow$ LLaVA	APGD	70.10 (-22)	25.02 (-72)	18.38 (-42)	8.64 (-73)
$\text{BLIP}_v \rightarrow$ LLaVA	CW	79.00 (-12)	72.54 (-20)	22.06 (-31)	19.92 (-37)
CLIP $\rightarrow$ BLIP2	PGD	76.48 (-9)	50.44 (-40)	29.66 (-17)	20.28 (-43)
CLIP $\rightarrow$ BLIP2	APGD	73.44 (-13)	36.18 (-57)	28.88 (-19)	16.40 (-54)
CLIP $\rightarrow$ BLIP2	CW	76.30 (-9)	72 (-14)	30.34 (-15)	28.58 (-20)
CLIP $\rightarrow$ Ins.BLIP	PGD	76.48 (-9)	50.44 (-40)	23.34 (-14)	16.58 (-39)
CLIP $\rightarrow$ Ins.BLIP	APGD	73.44 (-13)	36.18 (-57)	22.42 (-17)	12.94 (-52)
CLIP $\rightarrow$ Ins.BLIP	CW	76.30 (-9)	72 (-14)	24.00 (-11)	23.64 (-12)

Table 1. Classification acc.@1 under untargeted transfer attacks for COCO and Food-101 [1]. For brevity, BLIP_v refers to BLIP visual encoder, and Ins.BLIP refers to InstructBLIP. "Vis.Enc. Acc" shows the target LMMs' visual encoder's accuracy with adversarial input generated with the original visual encoder. Numbers in parenthesis show % change w.r.t. the pre-attack accuracy

Model	Attack	Visual Encoder Acc (%) Pre	Visual Encoder Acc (%) Post_N	Visual Encoder Acc (%) Post_S	LMM Acc (%) Pre	LMM Acc (%) Post_N	LMM Acc (%) Post_S
Food-101							
LLaVA1.5	PGD	90.18	11.21	0.44	31.80	5.00	1.18
LLaVA1.5	APGD	90.18	1.53	0.00	31.80	4.66	2.38
LLaVA1.5	CW	90.18	0.23	9.77	31.80	18.04	16.16
BLIP2 T5	PGD	84.07	0.61	0.02	35.79	1.14	0.14
BLIP2 T5	APGD	84.07	0.13	0.00	35.79	3.88	3.62
BLIP2 T5	CW	84.07	4.39	0.59	35.79	18.32	15.18
InstructBLIP	PGD	84.07	0.61	0.02	27.01	1.38	0.04
InstructBLIP	APGD	84.07	0.13	0.00	27.01	3.78	3.62
InstructBLIP	CW	84.07	4.39	0.59	27.01	19.90	16.62
Stanford Cars							
LLaVA1.5	PGD	77.42	3.01	0.01	37.62	5.94	0.87
LLaVA1.5	APGD	77.42	0.45	0.02	37.62	6.74	3.82
LLaVA1.5	CW	77.42	6.04	0.33	37.62	25.88	23.48
BLIP2 T5	PGD	79.23	0.07	0.00	61.75	0.52	0.54
BLIP2 T5	APGD	79.23	0.00	0.00	61.75	0.54	0.39
BLIP2 T5	CW	79.23	10.35	1.53	61.75	0.63	0.61
InstructBLIP	PGD	79.23	0.07	0.00	16.86	0.47	0.02
InstructBLIP	APGD	79.23	0.00	0.00	16.86	0.54	0.39
InstructBLIP	CW	79.23	10.35	1.53	16.86	13.43	12.59

Table 2. Top-1 image classification result on Food-101 and Stanford Cars [6]. The "Visual Encoder Acc (%)" column refers to each LMM's visual encoder's accuracy (CLIP for LLaVA1.5, EVA-CLIP for BLIP2 and InstructBLIP).

1. Attacks

We use implementations from torchattacks [5] to generate adversarial images. All the attacks are un-targeted. Below in Fig. 1, we show more visualizations of adversarial images under different attacks, attack strength and models.

Among three types of attacks, APGD [2] generates the most perceptible perturbations, as can be seen in Fig. 1. On the contrary, CW yields the most imperceptible perturbations, even under the strong setting. All the three attacks are imperceptible under the normal setting.

By comparing adversarial images generated by CLIP and EVA-CLIP (top and bottom row from Fig. 1), we can observe adversarial perturbations generated under EVA-CLIP are generally more perceptible than that of CLIP's. We can also observe the interestingly highlighted patch border for EVA-CLIP under APGD_S, which does not exist on that of CLIP's.

2. Additional experiments on classification

In Table 2 we present additional classification results for two fine-grained datasets: Food-101 and Stanford Cars. These results corroborate our prior observations, indicating that LMMs exhibit susceptibility to visual adversarial inputs, with their performance directly related to the robustness of their visual encoders.

3. Transferability of Visual Attacks on LMMs

In Table 1 we provide preliminary results on transferability of visual attacks against LMMs by applying perturbations generated for one visual encoder to a non-matching LMM. We observe that the impact of transfer attack on LLMs is directly linked to the impact of transfer attack on their corresponding visual encoders. Therefore, in terms of visual transfer attack, studying its effect on LMMs may simply be reduced back to studying its effect on standalone vision models.

4. LLM Responses to Adversarial Visual Questions

Fig. 2 shows more results on per-question accuracy drop after adversarial attack. We observe consistent behavior across three tested LMMs (LLaVA [9], BLIP2-T5 [7] and InstructBLIP [3], where accuracy drop the most on questions querying object types or attributes, such as "what animal/room/kind/type...".

In Fig. 3 and 4, we show more visualization on the three evaluated LMMs' responses to APGD and CW attacks under the strong setting. We can again observe that adversarial

images under APGD attack cause all three LMMs to output completely incorrect image descriptions, yet still having correct answers for "peripheral" questions, especially those querying the backgrounds. Notably, adversarial images generated under CW attack have little impact on all three LMMs, despite the relatively low classification accuracy after CW attack for CLIP and EVA-CLIP.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 1

[2] 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. 1

[3] 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

[4] 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, 5

[5] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020. 1

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[7] 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

[8] 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. 3

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

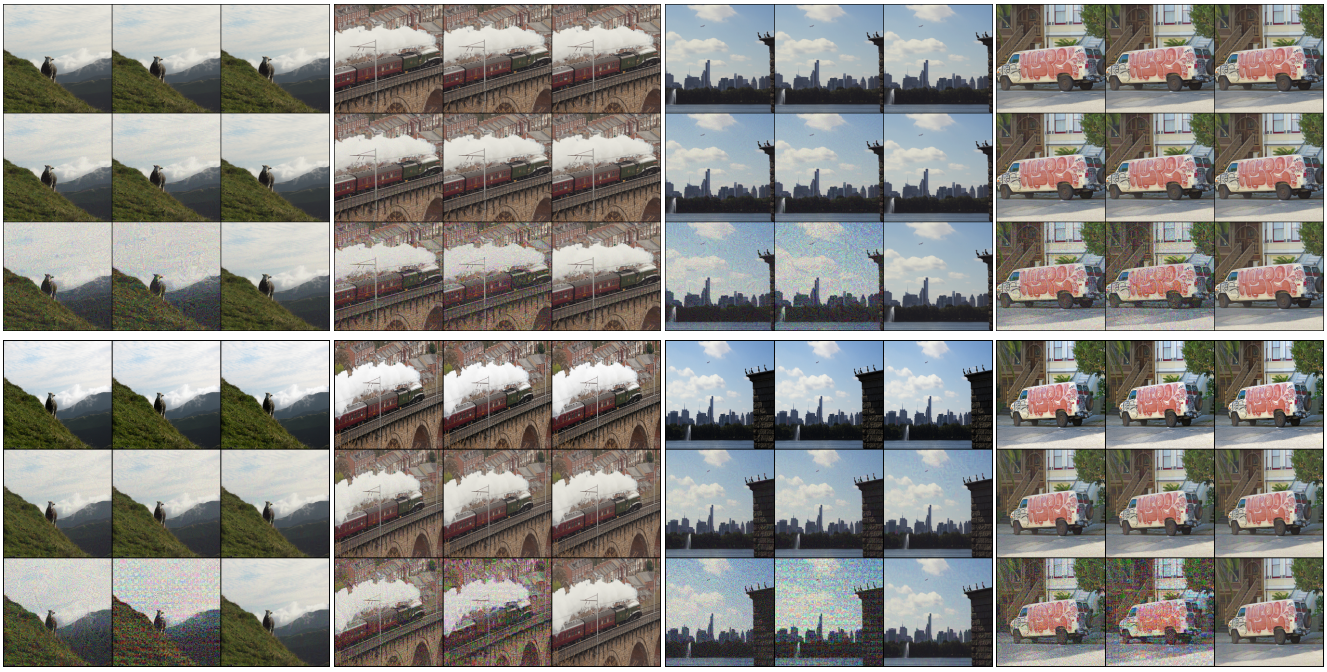


Figure 1. Visualization of three attacks generated from CLIP (top) and EVA-CLIP (bottom). In each of the 3×3 cell, top/mid/bottom row is from clean/normal/strong, and left/mid/right column is from PGD/APGD/CW attack, respectively. Image source: COCO val2014 [8].

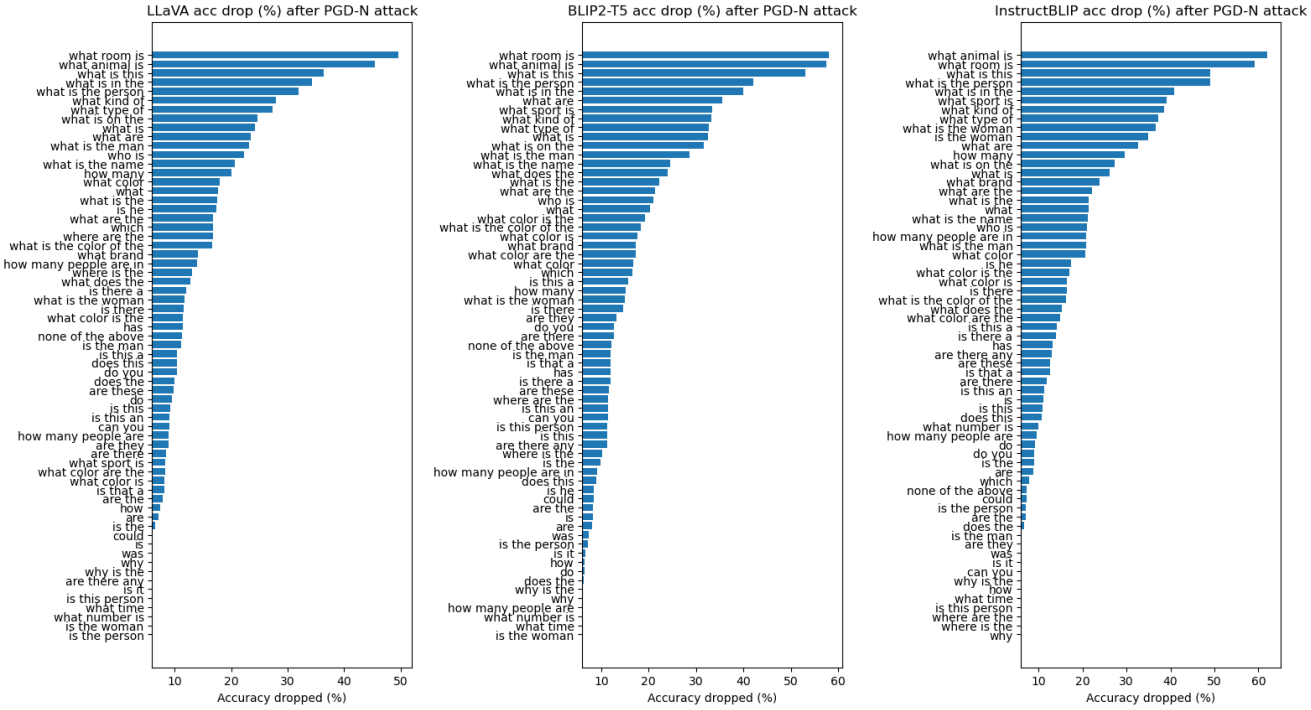


Figure 2. VQA V2 per question-type accuracy drop on LLaVA, BLIP2-T5 and InstructBLIP under PGD_N.

What is this image about?

LLaVA: A street sign for Broadway is hanging from a pole. ✓

LLaVA(APGD): A bowl filled with shrimp and other seafood. ✗

LLaVA(CW): A street sign for Broadway is hanging from a pole. ✓

Should a pedestrian crossing Broadway on 7th stop?

LLaVA(APGD): No. ✓

LLaVA(CW): Yes. ✓

What City is this taken in?

LLaVA(APGD): Nashville. ✗

LLaVA(CW): New york. ✓

What color is the writing on the front of the building?

LLaVA(APGD): White. ✓

LLaVA(CW): White. ✓

What is this image about?

InstructBLIP: a large billboard advertising jeans and outdoor clothing ✓

BLIP2-T5: a large advertisement for brooklyn eagle outfitters ✓

InstructBLIP(APGD): a bag of butter is sitting on top of a table ✗

BLIP2-T5(APGD): a bag of eagle butter ✗

InstructBLIP(CW): a picture of a street sign on top of a building ✓

BLIP2-T5(CW): a large advertisement on a building ✓

Should a pedestrian crossing Broadway on 7th stop?

InstructBLIP(APGD): yes ✓

BLIP2-T5(APGD): no ✓

InstructBLIP(CW): yes ✓

BLIP2-T5(CW): no ✓

What City is this taken in?

InstructBLIP(APGD): new york ✓

BLIP2-T5: new york city ✓

InstructBLIP(CW): new york ✓

BLIP2-T5(APGD): new york ✓

BLIP2-T5(CW): new york city ✓

What color is the writing on the front of the building?

InstructBLIP: red ✗

BLIP2-T5(APGD): blue ✓

InstructBLIP(APGD): blue ✓

BLIP2-T5(CW): blue ✓

InstructBLIP(CW): white ✓

What is this image about?

LLaVA: A pastry with a hole in the center is covered in white frosting. ✓

LLaVA(APGD): A bowl filled with shrimp and other seafood. ✗

LLaVA(CW): A pastry with a hole in the center is covered in frosting. ✓

What color is the food on?

LLaVA(APGD): White. ✓

LLaVA(CW): White. ✓

Is this a healthy meal?

LLaVA(APGD): Yes. ✗

LLaVA(CW): No. ✓

What is this food?

LLaVA(APGD): Shrimp. ✗

LLaVA(CW): Donut. ✓

What is this image about?

InstructBLIP: a doughnut with white frosting and a hole in the middle ✓

BLIP2-T5: a donut with icing on it is sitting on a piece of paper ✓

InstructBLIP(APGD): a vase is sitting on a table next to a pair of shoes ✗

BLIP2-T5(APGD): a vase on a table ✗

InstructBLIP(CW): a small cake with white frosting on top of it ✓

BLIP2-T5(CW): white icing on a pastry ✓

What color is the food on?

InstructBLIP(APGD): no food ✗

BLIP2-T5(APGD): green ✗

InstructBLIP(CW): white ✓

BLIP2-T5(CW): white ✓

Is this a healthy meal?

InstructBLIP(APGD): no ✓

BLIP2-T5: no it's a donut ✓

InstructBLIP(CW): no ✓

BLIP2-T5(APGD): no ✓

BLIP2-T5(CW): no ✓

What is this food?

InstructBLIP: donut ✓

BLIP2-T5: pastry ✓

InstructBLIP(APGD): no ✗

BLIP2-T5(APGD): rice ✗

InstructBLIP(CW): cake ✓

BLIP2-T5(CW): pastry ✓

What is this image about?

LLaVA: A white bird with a red eye is standing in front of a green tree. ✓

LLaVA(APGD): A painting of a forest with strange creatures and a rock with a face. ✗

LLaVA(CW): A white bird with a black beak is standing in front of a green tree. ✓

What color is the animal?

LLaVA(APGD): White. ✓

LLaVA(CW): White. ✓

How can you tell this is a tropical location?

LLaVA(APGD): Palm trees. ✓

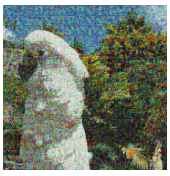
LLaVA(CW): Palm trees. ✓

What animal is this?

LLaVA: Parrot. ✓

LLaVA(APGD): Human. ✗

LLaVA(CW): Parrot. ✓



What is this image about?

InstructBLIP: a white cockatoo is standing in front of a palm tree ✓

BLIP2-T5: a statue of a white parrot ✓

InstructBLIP(APGD): an image of an orange spiral pattern on a white background ✗

BLIP2-T5(APGD): an orange background with a lot of circles ✗

InstructBLIP(CW): a large white statue of a parrot in a park ✓

BLIP2-T5(CW): a stuffed white cockatoo standing in a garden ✓

What color is the animal?

InstructBLIP(APGD): white ✓

BLIP2-T5(APGD): orange ✗

InstructBLIP(CW): white ✓

BLIP2-T5(CW): white ✓

How can you tell this is a tropical location?

InstructBLIP(APGD): palm trees ✓

BLIP2-T5(APGD): there are palm trees ✓

InstructBLIP(CW): palm trees ✓

BLIP2-T5(CW): there are palm trees ✓

What animal is this?

InstructBLIP: parrot ✓

BLIP2-T5: parrot ✓

InstructBLIP(APGD): none ✗

BLIP2-T5(APGD): zebra ✗

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

LLaVA(CW): A white bird with a black beak is standing in front of a green tree. ✓

How many statues can you see?

LLaVA: 5. ✓

LLaVA(APGD): 4. ✓

LLaVA(CW): 5. ✓

Is it night time?

LLaVA(APGD): No. ✓

LLaVA(CW): No. ✓

What time does the clock read?

LLaVA: 12:00. ✗

LLaVA(APGD): 12:00. ✗

LLaVA(CW): 12:00. ✗

What color is the clock face?

LLaVA(APGD): Blue. ✓

LLaVA(CW): Blue. ✓

What is this image about?

InstructBLIP: a clock on top of a building with statues on it ✓

BLIP2-T5: a clock on top of a building ✓

InstructBLIP(APGD): an image of a bird flying in front of a blue sky ✗

BLIP2-T5(APGD): a man is standing in the middle of a field ✗

InstructBLIP(CW): an image of a temple with statues on top of it ✓

BLIP2-T5(CW): the top of a temple with statues on it ✓

How many statues can you see?

InstructBLIP: 4 ✓

BLIP2-T5: three ✗

InstructBLIP(APGD): 1 ✓

BLIP2-T5(APGD): three ✗

InstructBLIP(CW): 1 ✓

BLIP2-T5(CW): three ✗

Is it night time?

InstructBLIP(APGD): no ✓

BLIP2-T5(APGD): no ✓

InstructBLIP(CW): no ✓

BLIP2-T5(CW): no ✓

What time does the clock read?

InstructBLIP: ✗

BLIP2-T5: 12 noon ✗

InstructBLIP(APGD): ✗

BLIP2-T5(APGD): 12 noon ✗

InstructBLIP(CW): ✗

BLIP2-T5(CW): 12 noon ✗

What color is the clock face?

InstructBLIP(APGD): white ✗

BLIP2-T5(APGD): blue ✓

InstructBLIP(CW): blue ✓

BLIP2-T5(CW): white ✗

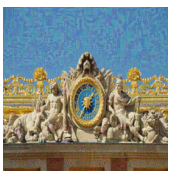


Figure 3. A comparison between the response from three LMMs (LLaVA, BLIP2-T5 and InstructBLIP) on adversarial image generated by APGD_S and CW_S with CLIP for LLaVA (left), and EVA-CLIP for BLIP2-T5 and InstructBLIP (mid and right). "(adv)" refers to LMM's response with the adversarial image. Within each cell, the top/bottom adversarial image is generated by APGD_S/CW_S, respectively. We show the clean response when the adversarial response is different. Image and questions source: VQA V2 [4].

What is this image about?
LLaVA: Two giraffes are standing in a field, one is taller and the other is shorter.✓
LLaVA(adv): A painting of a man and a dog pulling a cart with a crowd of people behind them. ✗

Is the giraffe a baby?
LLaVA: Yes ✓
LLaVA(adv): No ✓

How many giraffes are there?
LLaVA: 2 ✓
LLaVA(adv): 0 ✗

Are the animals in the shade?
LLaVA(adv): No ✓

How many giraffes can been seen?
LLaVA: 2 ✓
LLaVA(adv): 0 ✗

Are the animals in the wild?
LLaVA(adv): No ✓

Why is there a white roofed open pavilion?
LLaVA: Shade ✓
LLaVA(adv): No ✗

Will these giraffes be eating the grass for dinner?
LLaVA(adv): No ✓

Where do these animals live?
LLaVA: Zoo ✓
LLaVA(adv): Farm ✗

What are these animals?
LLaVA: Giraffes ✓
LLaVA(adv): Dogs ✗

Do you see a body of water in the picture?
LLaVA(adv): No ✓

What is in front of the giraffe?
LLaVA: Fence ✗
LLaVA(adv): People ✗

What is this image about?
BLIP2-T5: two giraffes standing in a field.✓
BLIP2-T5(adv): an abstract image of a circular pattern. ✗

Is the giraffe a baby?
BLIP2-T5: yes ✓
BLIP2-T5(adv): the giraffe is a baby ✓

How many giraffes are there?
BLIP2-T5: two ✓
BLIP2-T5(adv): there are two giraffes ✓

Are the animals in the shade?
BLIP2-T5: no ✓
BLIP2-T5(adv): are the animals in the shade? ✗

How many giraffes can been seen?
BLIP2-T5: two ✓
BLIP2-T5(adv): two ✓

Are the animals in the wild?
BLIP2-T5: no ✓
BLIP2-T5(adv): are the animals in the wild? ✗

Why is there a white roofed open pavilion?
BLIP2-T5: it is a zoo ✗
BLIP2-T5(adv): it's a relic from the ✗

Will these giraffes be eating the grass for dinner?
BLIP2-T5: yes ✓
BLIP2-T5(adv): no giraffes, gira ✗

Where do these animals live?
BLIP2-T5: zoo ✓
BLIP2-T5(adv): in the ocean ✗

What are these animals?
BLIP2-T5: giraffes ✓
BLIP2-T5(adv): octopus ✗

Do you see a body of water in the picture?
BLIP2-T5(adv): no ✓

What is in front of the giraffe?
BLIP2-T5: tree ✗
BLIP2-T5(adv): a rainbow ✗

What is this image about?
InstructBLIP: a couple of giraffes are standing in a dirt field.✓
InstructBLIP(adv): an abstract image of a colorful spiral pattern ✗

Is the giraffe a baby?
InstructBLIP: yes ✓
InstructBLIP(adv): no ✓

How many giraffes are there?
InstructBLIP: 2 ✓
InstructBLIP(adv): 0 ✗

Are the animals in the shade?
InstructBLIP(adv): no ✓

How many giraffes can been seen?
InstructBLIP: 2 ✓
InstructBLIP(adv): 0 ✗

Are the animals in the wild?
InstructBLIP(adv): no ✓

Why is there a white roofed open pavilion?
InstructBLIP(adv): yes ✗

Will these giraffes be eating the grass for dinner?
InstructBLIP(adv): yes ✓
InstructBLIP(adv): no ✓

Where do these animals live?
InstructBLIP: zoo ✓
InstructBLIP(adv): ocean ✗

What are these animals?
InstructBLIP: giraffes ✓
InstructBLIP(adv): none ✗

Do you see a body of water in the picture?
InstructBLIP(adv): no ✓

What is in front of the giraffe?
InstructBLIP(adv): nothing ✓

Figure 4. A comparison between responses from three LMMs (LLaVA, BLIP2-T5 and InstructBLIP) on adversarial images generated by $APGD_S$ with CLIP for LLaVA (left), and EVA-CLIP for BLIP2-T5 and InstructBLIP (mid and right). "(adv)" refers to LMM's response with the adversarial image. We show the clean response when the adversarial response is different. Image and questions source: VQA V2 [4].