

Appendix: Break the Visual Perception: Adversarial Attacks Targeting Encoded Visual Tokens of Large Vision-Language Models

Anonymous Authors

A VICTIM MODEL DETAILS

The details of victim large vision-language models are illustrated in Table 1. Various LVLs employ LLMs with large number of parameters, including LLaMA [9], MPT [8], OPT [11], and Vicuna [1]. To reduce randomness, we consistently employ greedy search to generate answers for clean images or adversarial images.

Table 1: The image encoders and base LLMs of victim models.

Victim Model	Image Encoder	Base LLM
LLaVA	OpenAI CLIP ViT-L	LLaMA-2-13b
Otter	OpenAI CLIP ViT-L	MPT-7b
LLaMA-Adapter-V2	OpenAI CLIP ViT-L	LLaMA-7b
OpenFlamingo	OpenAI CLIP ViT-L	MPT-7b
MiniGPT-4	EVA CLIP ViT-G	Vicuna-7b
BLIP-2	EVA CLIP ViT-G	OPT-2.7b
InstructBLIP	EVA CLIP ViT-G	Vicuna-7b
mPLUG-Owl-2	ViT-L (non-pretrained)	LLaMA-2-7b

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 VT-Attack Results against Different LVLs

We provide additional qualitative results of VT-Attack against different LVLs. Table 3, Table 4 and Table 5 present the results of attacking LLaVA [5], MiniGPT-4 [12] and mPLUG-Owl2 [10]. It can be observed that our proposed VT-Attack leads to outputs that are less relevant to the original answers compared to the baseline methods. This further illustrates the greater disruption caused by VT-Attack on information in visual tokens.

B.2 Results of Same Adversarial Image Attacking Various LVLs

Figure 1 and Figure 2 present the additional comparison of the answers generated by different LVLs when taking the same clean image and adversarial image as inputs. It can be observed that the same adversarial image leads to different outputs from different models. This indicates that the adversarial images whose encoded visual tokens are disrupted may not exhibit strong semantics; otherwise, different models should generate similar outputs.

B.3 Breakdown of LLaVA to Non-Visual Question Answering

In experiments we have observed that when adversarial images generated by VT-Attack are input to LLaVA [5], the question-answering capability of LLaVA not only exhibit failure to image-based queries, but also break in addressing non-visual prompts. Even when the

questions posed are unrelated to the images such as "What is artificial intelligence?", the model fails to provide reasonable responses. An example is depicted in Table 6. One possible reason is that the intermediate module of LLaVA is only a linear layer with a significantly smaller number of parameters. Therefore, LLaVA is more sensitive to corrupted visual tokens compared to other LVLs, even when queried with non-visual questions.

C ATTACK PERFORMANCE OVER ITERATIONS

The influence of attack iterations on attack performance (CLIP score) is illustrated in Figure 3, taking LLaVA [5] as an example. Increasing the number of attack iterations generally enhances the attack effectiveness. However, the improvement becomes less significant when the number of iterations exceeds 800.

D PROMPTS FOR DIFFERENT TASKS

To investigate the generality of adversarial examples across different prompts, we employ three question-answering tasks. Below are prompts used for image captioning task.

1. Describe this image briefly in one sentence.
2. Give a brief description of the image.
3. Can you provide a brief description of this image?
4. Offer a brief caption for this image.
5. Give a short overview of this image.
6. Provide a short summary of this picture.
7. Describe this picture in a few words.
8. Summarize this image briefly.
9. Please provide a short title for this image.
10. Briefly explain what is shown in this image.

The prompts for general VQA are illustrated below.

1. Is there a mobile phone in this image?
2. Is there any text or writing visible in the image?
3. Can you see any shoes in this image?
4. How many pens are visible in this picture?
5. Any signs of human activity in the image?
6. Where was this image taken?
7. Can you see animals in this image?
8. Can you identify any vehicles?
9. Are there any objects related to food?
10. Do you notice any body of water?

The prompts for detailed VQA are shown as below.

- 117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174
- 175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232
- 117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174
- 175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232
- 117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174
- 175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

E SENSITIVITY TO PIXEL NOISE DEFENSE

Adding tiny Gaussian noise to adversarial images is a simple way of model defense. To investigate the sensitivity of adversarial examples to it, we conduct experiments for BLIP-2 [4] as an example, and the results are shown in Table 2. Adversarial images are not sensitive to pixel noise if the size is relatively small. When the noise size exceeds 5/255, the attack performance starts to decline. However, overall, Gaussian noise within the same magnitude (8/255) as adversarial perturbations does not significantly affect the attack performance.

Table 2: The impact of adding Gaussian noise to adversarial images on attack performance (CLIP Score ↓) compared to clean images.

Noise Size	Clean	Adversarial
0/255	30.44	20.32
1/255	30.41	20.41
2/255	30.47	20.37
3/255	30.19	20.36
4/255	30.38	20.48
5/255	30.55	20.77
6/255	30.21	21.12
7/255	30.37	21.90
8/255	30.29	22.67

Table 3: Additional cases of attack results against LLaVA [5] in different methods.





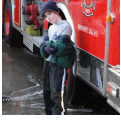





Image	Method	LVLM-Output	Image	Method	LVLM-Output
	No Attack	a group of monkeys on a tree		No Attack	a row of police cars parked on a city street
	E2E [7]	a tree filled with many birds		E2E [7]	two green police cars driving down a street
	CLIP-Based [2]	a black bear holding an object		CLIP-Based [2]	a clock on a building, a truck, and a person
	VT-Attack	a person holding a cup		VT-Attack	a white building with a clock
	No Attack	a white dog sitting on the grass		No Attack	a plaque on a brick wall
	E2E [7]	a large dog walking through the snow		E2E [7]	a close-up of a gravestone with a name on it
	CLIP-Based [2]	a dog with a collar and a tag		CLIP-Based [2]	a building with a sign on it
	VT-Attack	the a question, the the, the, the		VT-Attack	a a, a, a, a a, a a, a a a in the
	No Attack	a young boy standing next to a fire truck		No Attack	a hummingbird perched on a thin branch
	E2E [7]	an elderly woman walking down the street		E2E [7]	a bird standing on a bamboo stick
	CLIP-Based [2]	two women posing for a picture in a park		CLIP-Based [2]	a bird with a long beak on a wooden pole
	VT-Attack	a man is holding a red cell phone		VT-Attack	the question, the question, the, the question
	No Attack	a pile of yarn sitting on a table		No Attack	a small brown and white dog lying on the grass
	E2E [7]	a crochet hook and a ball of yarn		E2E [7]	a large brown dog playing with toy horses
	CLIP-Based [2]	a pair of old, worn-out leather boots		CLIP-Based [2]	a large brown bear with paws
	VT-Attack	a yellow cloth bag with a yellow cloth		VT-Attack	a person is seen holding a piece of wood
	No Attack	a bottle of Fanta and a can of Coca-Cola		No Attack	a pair of Nike sneakers sitting in a box
	E2E [7]	a bottle of green sauce or dressing		E2E [7]	a woman holding a large pile of shoes
	CLIP-Based [2]	a person holding a box of tissues		CLIP-Based [2]	a pair of shoes with a shredded shoe lace
	VT-Attack	a person is sitting at a table		VT-Attack	the the, the, the, the, the, the, the

Table 4: Additional cases of attack results against MiniGPT-4 [12] in different methods.


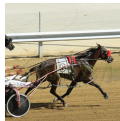




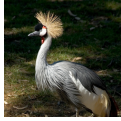


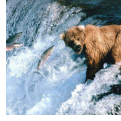
Image	Method	LVLM-Output	Image	Method	LVLM-Output
	No Attack	a man and a woman in a sailboat		No Attack	a carriage being pulled by horses
	E2E [7]	two people are sitting on a sailboat		E2E [7]	a carriage is being pulled by a man
	CLIP-Based [2]	a group of people on a boat in the river		CLIP-Based [2]	a man in a suit carrying a baby in a stroller
	VT-Attack	a dog sitting on a bench		VT-Attack	a small dog pulling a toy cart
	No Attack	a large ship with waves crashing		No Attack	a group of men playing basketball
	E2E [7]	a large cruise ship sailing on the ocean		E2E [7]	two men in basketball uniforms
	CLIP-Based [2]	a book cover with a picture of a boat		CLIP-Based [2]	a group of people playing basketball game
	VT-Attack	a close-up view of a compass with a needle		VT-Attack	a man standing in front of a vintage jukebox
	No Attack	a set of two padlocks with keys		No Attack	a white box with two green lights on top
	E2E [7]	a lock on a wooden table		E2E [7]	a pile of old electronics
	CLIP-Based [2]	a set of four metal bells		CLIP-Based [2]	a plastic box with a clear plastic
	VT-Attack	a vintage jukebox with many records		VT-Attack	2003 Toyota Matrix in 3D red
	No Attack	a bird with a long, curved beak		No Attack	a man holding a bicycle on his head
	E2E [7]	two birds standing on a dirt road		E2E [7]	a young boy holding a bicycle at sunset
	CLIP-Based [2]	a group of birds sitting on the ground		CLIP-Based [2]	a person carrying a bicycle on their back
	VT-Attack	a vintage music box with a colorful design		VT-Attack	a person riding a bicycle on a path
	No Attack	a hot pot of food on a stove		No Attack	a brown bear standing on a rocky cliff
	E2E [7]	a large pot filled with food		E2E [7]	a brown bear carrying a stuffed animal
	CLIP-Based [2]	a bowl of mixed seafood		CLIP-Based [2]	a close-up view of a large clamshell on a beach
	VT-Attack	a plate of crawfish etouffee		VT-Attack	a dog with a carrot in its mouth

Table 5: Additional cases of attack results against mPLUG-Owl-2 [10] in different methods.












Image	Method	LVLM-Output	Image	Method	LVLM-Output
	No Attack	a man with dreadlocks plays a trumpet		No Attack	a bird sitting on a branch in the sky
	E2E [7]	a girl is playing a trumpet		E2E [7]	a man holding a baby bird in his hand
	VT-Attack	a doll is shown in two different pictures		VT-Attack	a group of people standing together
	No Attack	a bridge over a river with trees		No Attack	a brown dog panting on a sidewalk
	E2E [7]	aerial view of san francisco bridge		E2E [7]	an ostrich wearing a hat
	VT-Attack	a close up of a pair of pants		VT-Attack	a picture of a bridge with a blue sky
	No Attack	a white and gray dog standing on grass		No Attack	a sewer cover with a metal design on it
	E2E [7]	a baby is sitting on top of a lion		E2E [7]	a castle is shown on a tire
	VT-Attack	a group of people are standing in a room		VT-Attack	two microphones are set up in a studio
	No Attack	a large building with a curved roof		No Attack	a bridge over a river with a train on it
	E2E [7]	a cruise ship in the ocean		E2E [7]	a man driving a car under a bridge
	VT-Attack	a pair of shoes with the number 43 on them		VT-Attack	a group of 777 nail polish bottles
	No Attack	a tank with people on top of it		No Attack	a baby wearing a green knitted hat
	E2E [7]	a dog driving a tank with a missile		E2E [7]	a man with a bubble on his face
	VT-Attack	a photo of a person in a black dress		VT-Attack	a close-up of a gold and black watch

Table 6: Breakdown of LLaVA to non-visual prompts when queried with adversarial image in VT-Attack.

Adversarial image	Non-visual Prompt	LLaVA-Output
	What is artificial intelligence?	in the a, a in the a, a in the a in the a in
	What is the capital of France?	a person is holding a red glass ball
	How do you bake a cake?	„„„, in the a„„„, in the a, a in the a in
	Why do leaves change color in the fall?	the question, the question, the question

Clean image	LLaVA	a man playing a trumpet on stage	OpenAI CLIP adv	LLaVA	... in the a,..., in the a
	Otter	a man plays the French horn on stage		Otter	a photo of a room with a person
→	LLaMA Adapter-v2	a man is playing a trumpet	→	LLaMA Adapter-v2	two men are sitting on a couch
	Open Flamingo	a saxophonist plays a solo		Open Flamingo	a man is holding a book
Clean image	LLaVA	a white ironing board with a blue cover	OpenAI CLIP adv	LLaVA	a in a a in a a in a a in a a
	Otter	an iron with a blue handle		Otter	a photo of a room with a window
→	LLaMA Adapter-v2	a white iron with a wooden handle	→	LLaMA Adapter-v2	a woman is watching TV
	Open Flamingo	a white iron with a blue handle		Open Flamingo	a man is sitting on a bench
Clean image	LLaVA	a shopping cart with an American flag	OpenAI CLIP adv	LLaVA	the, the, the, the, the, the, the, the
	Otter	a shopping cart with a large flag		Otter	the person is standing behind a screen
→	LLaMA Adapter-v2	a shopping cart with an American flag	→	LLaMA Adapter-v2	a black image of a newspaper
	Open Flamingo	a shopping cart in a parking garage		Open Flamingo	a man is standing in a doorway
Clean image	LLaVA	a yellow school bus on a wet street	OpenAI CLIP adv	LLaVA	a cat sitting on a truck's windshield
	Otter	a school bus is parked in a lot		Otter	a cat hanging off of a vehicle
→	LLaMA Adapter-v2	a school bus driving down a wet road	→	LLaMA Adapter-v2	a blue truck with an advertisement
	Open Flamingo	a yellow school bus		Open Flamingo	a close up of a bottle of beer
Clean image	LLaVA	a group of mushrooms in a forest	OpenAI CLIP adv	LLaVA	a person holding a bunch of white balls
	Otter	a bunch of mushrooms are growing		Otter	a man is standing near umbrellas
→	LLaMA Adapter-v2	a group of mushrooms growing together	→	LLaMA Adapter-v2	several jellyfish floating in the water
	Open Flamingo	a group of mushrooms in the forest		Open Flamingo	a man in a black shirt and black pants
Clean image	LLaVA	a large, open, and ornate shopping mall	OpenAI CLIP adv	LLaVA	a person in a room with a large screen
	Otter	an escalator in a mall with people on it		Otter	a blurry image of a building
→	LLaMA Adapter-v2	a large, open, and well-lit shopping mall	→	LLaMA Adapter-v2	a group of people standing in pool
	Open Flamingo	the escalator at the Venetian Hotel		Open Flamingo	an interior of Palais Garnier opera house
Clean image	LLaVA	a DJ turntable with a black color	OpenAI CLIP adv	LLaVA	a glass bowl
	Otter	a close up view of a sound mixer		Otter	a dark room with a light shining on it
→	LLaMA Adapter-v2	a large DJ turntable with a Pioneer logo	→	LLaMA Adapter-v2	a man is holding a cell phone
	Open Flamingo	the Pioneer CDJ-2000 Nexus		Open Flamingo	a close up of a woman's face

Figure 1: Additional comparison of the responses generated by various LVLMS when queried with clean images and adversarial images in VT-Attack against OpenAI CLIP [6].






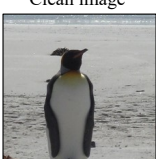


Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>a red hoodie with a logo in the middle</td></tr><tr><td>MiniGPT-4</td><td>a red hoodie with a word "georgia" on it</td></tr><tr><td>InstructBLIP</td><td>a red hoodie with the word "Georgia"</td></tr></table>	BLIP-2	a red hoodie with a logo in the middle	MiniGPT-4	a red hoodie with a word "georgia" on it	InstructBLIP	a red hoodie with the word "Georgia"	→	<table><tr><td>BLIP-2</td><td>a photo of a room with a plant</td></tr><tr><td>MiniGPT-4</td><td>a person holding a cell phone</td></tr><tr><td>InstructBLIP</td><td>a vintage image of a gas pump</td></tr></table>	BLIP-2	a photo of a room with a plant	MiniGPT-4	a person holding a cell phone	InstructBLIP	a vintage image of a gas pump
BLIP-2	a red hoodie with a logo in the middle															
MiniGPT-4	a red hoodie with a word "georgia" on it															
InstructBLIP	a red hoodie with the word "Georgia"															
BLIP-2	a photo of a room with a plant															
MiniGPT-4	a person holding a cell phone															
InstructBLIP	a vintage image of a gas pump															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>a blue and white pillow</td></tr><tr><td>MiniGPT-4</td><td>a green and brown leafy pattern</td></tr><tr><td>InstructBLIP</td><td>a blue and yellow pillow on a bed</td></tr></table>	BLIP-2	a blue and white pillow	MiniGPT-4	a green and brown leafy pattern	InstructBLIP	a blue and yellow pillow on a bed	→	<table><tr><td>BLIP-2</td><td>a map of Hawaii with a flower on it</td></tr><tr><td>MiniGPT-4</td><td>a close-up of a coffee cup</td></tr><tr><td>InstructBLIP</td><td>a black car is parked on a street</td></tr></table>	BLIP-2	a map of Hawaii with a flower on it	MiniGPT-4	a close-up of a coffee cup	InstructBLIP	a black car is parked on a street
BLIP-2	a blue and white pillow															
MiniGPT-4	a green and brown leafy pattern															
InstructBLIP	a blue and yellow pillow on a bed															
BLIP-2	a map of Hawaii with a flower on it															
MiniGPT-4	a close-up of a coffee cup															
InstructBLIP	a black car is parked on a street															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>a stop sign and two street signs on a pole</td></tr><tr><td>MiniGPT-4</td><td>a street sign with a stop sign</td></tr><tr><td>InstructBLIP</td><td>a street sign with the name "McLean"</td></tr></table>	BLIP-2	a stop sign and two street signs on a pole	MiniGPT-4	a street sign with a stop sign	InstructBLIP	a street sign with the name "McLean"	→	<table><tr><td>BLIP-2</td><td>a plant with a person sitting next to it</td></tr><tr><td>MiniGPT-4</td><td>a woman is holding a glass of wine</td></tr><tr><td>InstructBLIP</td><td>a man is playing a trumpet</td></tr></table>	BLIP-2	a plant with a person sitting next to it	MiniGPT-4	a woman is holding a glass of wine	InstructBLIP	a man is playing a trumpet
BLIP-2	a stop sign and two street signs on a pole															
MiniGPT-4	a street sign with a stop sign															
InstructBLIP	a street sign with the name "McLean"															
BLIP-2	a plant with a person sitting next to it															
MiniGPT-4	a woman is holding a glass of wine															
InstructBLIP	a man is playing a trumpet															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>a pile of plastic pens</td></tr><tr><td>MiniGPT-4</td><td>a large pile of colorful pencils</td></tr><tr><td>InstructBLIP</td><td>a close-up view of a pile of pencils</td></tr></table>	BLIP-2	a pile of plastic pens	MiniGPT-4	a large pile of colorful pencils	InstructBLIP	a close-up view of a pile of pencils	→	<table><tr><td>BLIP-2</td><td>a close-up of a red machine tool</td></tr><tr><td>MiniGPT-4</td><td>a red and black object with a frame</td></tr><tr><td>InstructBLIP</td><td>a blue and purple bicycle</td></tr></table>	BLIP-2	a close-up of a red machine tool	MiniGPT-4	a red and black object with a frame	InstructBLIP	a blue and purple bicycle
BLIP-2	a pile of plastic pens															
MiniGPT-4	a large pile of colorful pencils															
InstructBLIP	a close-up view of a pile of pencils															
BLIP-2	a close-up of a red machine tool															
MiniGPT-4	a red and black object with a frame															
InstructBLIP	a blue and purple bicycle															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>the courthouse in the town</td></tr><tr><td>MiniGPT-4</td><td>a tall brick building with a clock tower</td></tr><tr><td>InstructBLIP</td><td>a large, old building with a clock tower</td></tr></table>	BLIP-2	the courthouse in the town	MiniGPT-4	a tall brick building with a clock tower	InstructBLIP	a large, old building with a clock tower	→	<table><tr><td>BLIP-2</td><td>a man in a suit and tie</td></tr><tr><td>MiniGPT-4</td><td>a glass bottle in the image</td></tr><tr><td>InstructBLIP</td><td>a man is walking down a street</td></tr></table>	BLIP-2	a man in a suit and tie	MiniGPT-4	a glass bottle in the image	InstructBLIP	a man is walking down a street
BLIP-2	the courthouse in the town															
MiniGPT-4	a tall brick building with a clock tower															
InstructBLIP	a large, old building with a clock tower															
BLIP-2	a man in a suit and tie															
MiniGPT-4	a glass bottle in the image															
InstructBLIP	a man is walking down a street															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>the king penguin is standing by himself</td></tr><tr><td>MiniGPT-4</td><td>a penguin standing on a beach</td></tr><tr><td>InstructBLIP</td><td>a penguin standing on a sandy beach</td></tr></table>	BLIP-2	the king penguin is standing by himself	MiniGPT-4	a penguin standing on a beach	InstructBLIP	a penguin standing on a sandy beach	→	<table><tr><td>BLIP-2</td><td>a blue swimming pool</td></tr><tr><td>MiniGPT-4</td><td>a close up of a sign</td></tr><tr><td>InstructBLIP</td><td>a man is playing a trumpet</td></tr></table>	BLIP-2	a blue swimming pool	MiniGPT-4	a close up of a sign	InstructBLIP	a man is playing a trumpet
BLIP-2	the king penguin is standing by himself															
MiniGPT-4	a penguin standing on a beach															
InstructBLIP	a penguin standing on a sandy beach															
BLIP-2	a blue swimming pool															
MiniGPT-4	a close up of a sign															
InstructBLIP	a man is playing a trumpet															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>a dining table and chairs</td></tr><tr><td>MiniGPT-4</td><td>a large wooden dining table with chairs</td></tr><tr><td>InstructBLIP</td><td>a dining table with a set of chairs</td></tr></table>	BLIP-2	a dining table and chairs	MiniGPT-4	a large wooden dining table with chairs	InstructBLIP	a dining table with a set of chairs	→	<table><tr><td>BLIP-2</td><td>a black and white photo of a room</td></tr><tr><td>MiniGPT-4</td><td>a dog and a cat are playing with a bird</td></tr><tr><td>InstructBLIP</td><td>a man is watching TV</td></tr></table>	BLIP-2	a black and white photo of a room	MiniGPT-4	a dog and a cat are playing with a bird	InstructBLIP	a man is watching TV
BLIP-2	a dining table and chairs															
MiniGPT-4	a large wooden dining table with chairs															
InstructBLIP	a dining table with a set of chairs															
BLIP-2	a black and white photo of a room															
MiniGPT-4	a dog and a cat are playing with a bird															
InstructBLIP	a man is watching TV															
Clean image			EVA CLIP adv													
	→	<table><tr><td>BLIP-2</td><td>a baseball player in uniform with a glove</td></tr><tr><td>MiniGPT-4</td><td>a baseball player in a uniform</td></tr><tr><td>InstructBLIP</td><td>a baseball player, likely a catcher</td></tr></table>	BLIP-2	a baseball player in uniform with a glove	MiniGPT-4	a baseball player in a uniform	InstructBLIP	a baseball player, likely a catcher	→	<table><tr><td>BLIP-2</td><td>the people are facing the camera</td></tr><tr><td>MiniGPT-4</td><td>a close-up of a man's face</td></tr><tr><td>InstructBLIP</td><td>a man and a woman are sitting together</td></tr></table>	BLIP-2	the people are facing the camera	MiniGPT-4	a close-up of a man's face	InstructBLIP	a man and a woman are sitting together
BLIP-2	a baseball player in uniform with a glove															
MiniGPT-4	a baseball player in a uniform															
InstructBLIP	a baseball player, likely a catcher															
BLIP-2	the people are facing the camera															
MiniGPT-4	a close-up of a man's face															
InstructBLIP	a man and a woman are sitting together															

Figure 2: Additional comparison of the responses generated by various LVLMs when queried with clean images and adversarial images in VT-Attack against EVA CLIP [3].

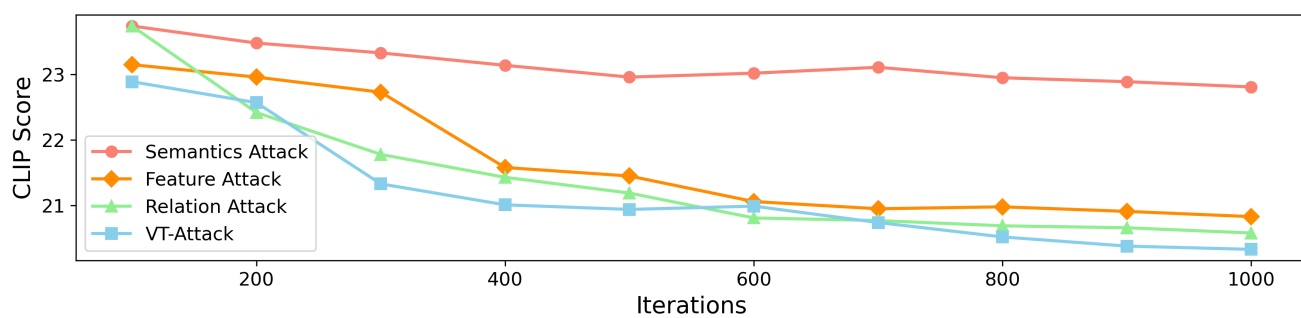


Figure 3: The relationship between attack performance and iterations (taking LLaVA [5] as an example).

REFERENCES

- [1] 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. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. <https://lmsys.org/blog/2023-03-30-vicuna/>
- [2] Xuanimng Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2023. On the Robustness of Large Multimodal Models Against Image Adversarial Attacks. *arXiv preprint arXiv:2312.03777* (2023).
- [3] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 19358–19369. <https://doi.org/10.1109/CVPR52729.2023.01855>
- [4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html>
- [5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *CoRR* abs/2304.08485 (2023). <https://doi.org/10.48550/arXiv.2304.08485>
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [7] Christian Schlarman and Matthias Hein. 2023. On the Adversarial Robustness of Multi-Modal Foundation Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*. IEEE, 3679–3687. <https://doi.org/10.1109/ICCVW60793.2023.00395>
- [8] MosaicML NLP Team. 2023. *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs*. www.mosaicml.com/blog/mpt-7b Accessed: 2023-05-05.
- [9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971* [cs.CL]
- [10] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257* (2023).
- [11] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv:2205.01068* [cs.CL]
- [12] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *CoRR* abs/2304.10592 (2023). <https://doi.org/10.48550/arXiv.2304.10592>