ADVERSARIAL ROBUSTNESS FOR VISUAL GROUNDING OF MULTIMODAL LARGE LANGUAGE MODELS

Kuofeng Gao[1], **Yang Bai**[2], **Jiawang Bai**[1], **Yong Yang**[2]†, **Shu-Tao Xia**[1,3]†
[1] Tsinghua University [2] Tencent Security Platform [3] Peng Cheng Laboratory
{gkf21,bjw19}@mails.tsinghua.edu.cn, mavisbai@tencent.com
coolcyang@tencent.com, xiast@sz.tsinghua.edu.cn

ABSTRACT

Multi-modal Large Language Models (MLLMs) have recently achieved enhanced performance across various vision-language tasks including visual grounding capabilities. However, the adversarial robustness of visual grounding remains unexplored in MLLMs. To fill this gap, we use referring expression comprehension (REC) as an example task in visual grounding and propose three adversarial attack paradigms as follows. Firstly, untargeted adversarial attacks induce MLLMs to generate incorrect bounding boxes for each object. Besides, exclusive targeted adversarial attacks cause all generated outputs to the same target bounding box. In addition, permuted targeted adversarial attacks aim to permute all bounding boxes among different objects within a single image. Extensive experiments demonstrate that the proposed methods can successfully attack visual grounding capabilities of MLLMs. Our methods not only provide a new perspective for designing novel attacks but also serve as a strong baseline for improving the adversarial robustness for visual grounding of MLLMs.

1 INTRODUCTION

Multi-modal Large Language Models (MLLMs) (Alayrac et al., 2022; Chen et al., 2022; Liu et al., 2023; Li et al., 2021; 2023), such as GPT-4 (OpenAI, 2023), integrate visual modality into large language models (LLMs) and have achieved state-of-the-art performance across various multi-modal tasks, including image captioning and visual question answering. Recent advancements in research (Chen et al., 2023a;b; Peng et al., 2023) have further unlocked the potential visual grounding capabilities of MLLMs. Through this grounding capability, MLLMs can accurately recognize objects, locate them, and provide visual responses, such as bounding boxes, thereby facilitating additional vision-language tasks, including referring expression comprehension.

Despite the impressive multi-modal performance of MLLMs, recent studies (Dong et al., 2023; Zhao et al., 2023; Carlini et al., 2023; Qi et al., 2023; Gao et al., 2024a;b; Yang et al., 2024) have revealed their susceptibility of MLLMs against adversarial attacks. Adversarial attacks manipulate input data with an imperceptible perturbation with the intention of misleading the model, often resulting in incorrect outputs. Most existing adversarial attacks on MLLMs have made main efforts on the image captioning and visual question answering task. Specifically, they craft an adversarial image that closely resembles the original image and employ it to prompt MLLMs, which can induce MLLMs to generate a wrong caption or reply an incorrect answer. However, the adversarial robustness on visual grounding is still unclear.

In this paper, we study the impact of adversarial attacks on visual grounding capabilities of MLLMs at first. As a representative example, we evaluate the adversarial robustness for visual grounding of MLLMs specifically through the task of referring expression comprehension. Referring expression comprehension (REC) is the process of identifying and localizing objects within an image based on a given textual prompt, ultimately generating bounding boxes of objects. Following previous work (Dong et al., 2023; Zhao et al., 2023), we focus on visual modality and aim to craft adversarial images with an imperceptible perturbation to perform adversarial attacks.

†Corresponding authors

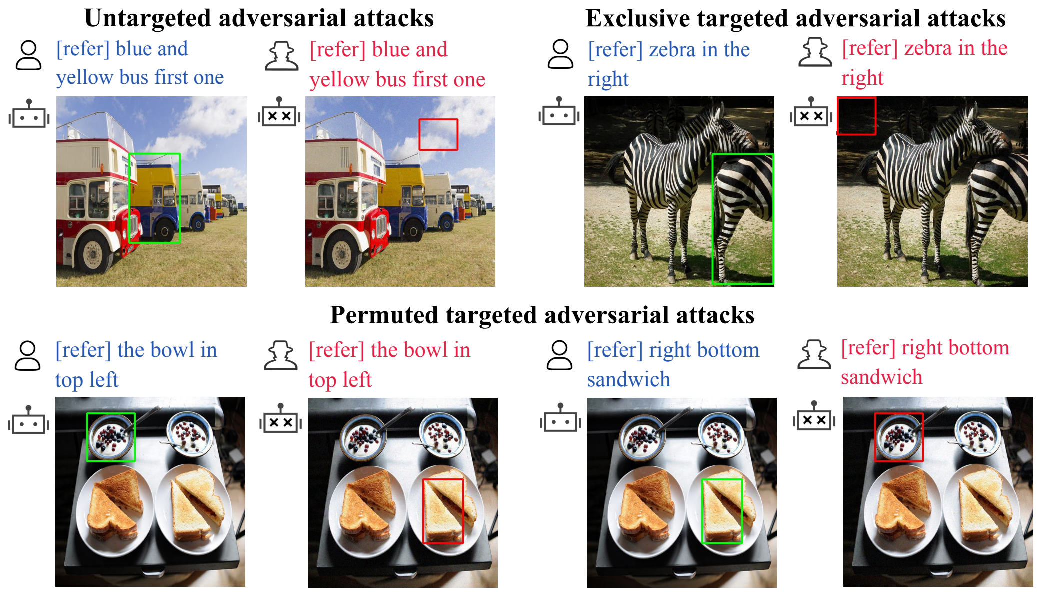

Figure 1: Three adversarial attack paradigms are proposed to evaluate the adversarial robustness for visual grounding of MLLMs.

Concretely, three attack paradigms are proposed tailored for REC of MLLMs as follows. Firstly, an untargeted attack aims to reduce the accuracy of bounding box predictions. This attack is deemed successful if the objects in the adversarial images are incorrectly located based on the original textual prompt. Besides, based on the type of target bounding box, two categories of targeted adversarial attacks are proposed, *i.e.*, exclusive targeted adversarial attacks and permuted targeted adversarial attacks. Exclusive targeted adversarial attacks deceive MLLMs to generate the same target bounding box, such as top left corner, regardless of their ground-truths. In contrast, permuted targeted adversarial attacks assign different target bounding boxes to different objects with the attacking goal of rearranging all bounding boxes within a single image.

The main contributions of this work are three-fold: (1) To the best of our knowledge, we are the first to reveal the adversarial threat in visual grounding of MLLMs. (2) We propose three attack paradigms to evaluate grounding adversarial robustness of MLLMs, including untargeted adversarial attacks, exclusive targeted adversarial attacks and permuted targeted adversarial attacks. (3) Extensive experiments are conducted, which verify the effectiveness of our proposed attacks.

2 The Proposed Attack

2.1 Preliminaries

Given an image $\boldsymbol{x}$ and multiple input textual prompts $T = \{t_i\}_{i=1}^N$, referring expression comprehension (REC) aims to locate corresponding target objects by bounding boxes $B = \{b_i\}_{i=1}^N$. During training of MLLMs, these bounding boxes are transformed into the textual formatting and MLLMs are trained using the auto-regressive loss. Successful REC by MLLMs is assumed when Intersection over Union (IoU) between the ground-truth and predicted bounding boxes exceeds 0.5.

Threat model. The goal of attackers is to optimize an imperceptible perturbation to craft adversarial images $\hat{\boldsymbol{x}}$ to achieve adversarial attacks. Specifically, the involved perturbation is restricted within a predefined magnitude ϵ in l_∞ norm, ensuring it difficult to detect. As suggested in Bagdasaryan et al. (2023); Qi et al. (2023), we assume that the victim MLLMs can be accessed in full knowledge, including both architectures and parameters of victim MLLMs.

2.2 Untargeted Adversarial Attacks

Untargeted adversarial attacks craft adversarial images $\hat{\boldsymbol{x}}$ with the aim of causing MLLMs to predict a bounding box that deviates from its ground-truth b_i when given an input textual prompt t_i. To this end, we propose two methods to mislead the MLLM's predictions, *i.e.*, **image embedding attack** and **textual bounding box attack**.

Image embedding attack. MLLMs first use vision encoders $f(\cdot)$ to extract image embeddings and generate the textual formatting of bounding boxes. Hence, image embedding attack can be implemented by maximizing the l_2 distance of the image embeddings between the original image $\boldsymbol{x}$ and the adversarial image $\hat{\boldsymbol{x}}$. The disrupted image embeddings will result in the model's inability to accurately predict the bounding boxes based on the input textual prompts. The objective function can be formulated as:

$$\max_{\boldsymbol{x}} ||f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x})||_2^2, \quad \text{s.t. } ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon. \tag{1}$$

Textual bounding box attack. Based on the original image $\boldsymbol{x}$ and the input textual prompt t_i, MLLMs $g(\cdot)$ will generate the textual formatting of the bounding box b_i in an auto-regressive manner. Concretely, MLLMs aim to estimate the probability of a next token given its context, including the original image $\boldsymbol{x}$, the input textual prompt t_i, and previous generated M tokens. Given the textual formatting of ground-truth bounding box $b_i = \{b_i^j\}_{j=1}^L$, the objective function can be formulated as:

$$\min_{\boldsymbol{x}} \sum_{j=1}^{L} \log p_g(b_i^j \mid b_i^{j<M};\ \boldsymbol{x};\ t_i), \quad \text{s.t. } ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon, \tag{2}$$

where $b_i^{j<M}$ denotes the previous generated M tokens. Textual bounding box attacks minimize the log-likelihood of the textual formatting of the ground-truth bounding box.

2.3 Targeted Adversarial Attacks

Targeted adversarial attacks craft adversarial images $\hat{\boldsymbol{x}}$ with the goal of causing MLLMs to predict a target bounding box different from the ground-truth bounding box b_i when given an input textual prompt t_i. Based on the type of target bounding box, two targeted attack paradigms are proposed, including **exclusive targeted adversarial attacks** and **permuted targeted adversarial attacks**.

Exclusive targeted adversarial attacks. Regardless of the input textual prompt, exclusive targeted adversarial attacks deceive MLLMs to locate all objects in images to the same target bounding box, denoted as b_u. To achieve this attack, given the textual formatting of target bounding box $b_u = \{b_u^j\}_{j=1}^L$, the objective function can be formulated as:

$$\max_{\boldsymbol{x}} \sum_{j=1}^{L} \log p_g(b_u^j \mid b_u^{j<M};\ \boldsymbol{x};\ t_i), \quad \text{s.t. } ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon, \tag{3}$$

where $b_u^{j<M}$ denotes previous generated M tokens. Exclusive targeted adversarial attacks maximize the log-likelihood of the textual formatting of the same target bounding box.

Permuted targeted adversarial attacks. Permuted targeted adversarial attacks aim to rearrange bounding box of all objects within an image. The target bounding box is determined based on the ground-truth bounding box. Given an input textual prompt t_i associated with the corresponding bounding box b_i, permuted targeted adversarial attacks set the target bounding box as $b_{(i+1) \bmod N}$, where N represents the number of objects within the image. This approach ensures that each object's bounding box is shifted to the next object, effectively rearranging all bounding boxes in the image. The objective function can be formulated as:

$$\max_{\boldsymbol{x}} \sum_{j=1}^{L} \log p_g(b_{(i+1) \bmod N}^j \mid b_{(i+1) \bmod N}^{j<M};\ \boldsymbol{x};\ t_i), \quad \text{s.t. } ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon, \tag{4}$$

where L denotes the token number of textual formatting of target bounding box and $b_{(i+1) \bmod N}^{j<M}$ denotes previous generated M tokens. Permuted targeted adversarial attacks maximize the log-likelihood of the textual formatting of the target bounding box, which is shifted from another object within an image.

Table 1: The IoU@0.5 (%) of two proposed untargeted adversarial attack methods against MiniGPT-v2 on three datasets. The lower values correspond to a stronger attack.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	test-A	test-B	val	test-A	test-B	val	test
No attack	84.96	89.39	82.15	76.22	82.57	70.30	81.61	82.01
Image embedding attacks	29.58	35.60	19.23	21.86	27.78	12.64	19.28	19.91
Textual bounding box attacks	43.60	49.60	36.58	36.18	42.42	28.65	36.74	37.41

Table 2: The IoU@0.5 (%) of two proposed targeted adversarial attack paradigms against MiniGPT-v2 on three datasets. The higher values correspond to a stronger attack.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	test-A	test-B	val	test-A	test-B	val	test
Exclusive (No attack)	0.14	0.08	0.22	0.11	0.05	0.21	0.20	0.04
Exclusive	62.12	63.94	60.98	61.93	62.90	61.11	60.96	60.77
Permuted (No attack)	5.69	5.17	7.43	10.65	7.87	14.1	10.09	10.15
Permuted	27.87	30.26	29.37	29.91	30.66	33.22	30.12	29.69

3 Experiments

3.1 Experimental Setups

Models and datasets. We consider the 7B version of MiniGPT-v2 Chen et al. (2023a) as the sandbox to launch our attack. Moreover, RefCOCO (Kazemzadeh et al., 2014) and RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) are considered as benchmark datasets for evaluation.

Baselines and setups. To optimize three proposed adversarial attacks, we perform the projected gradient descent (PGD) (Madry et al., 2018) algorithm in $T = 100$ iterations. Besides, the perturbation magnitude is set as $\epsilon = 16$ within l_∞ restriction, following Dong et al. (2023); Qi et al. (2023), and the step size is set as $\alpha = 1$. In exclusive targeted adversarial attacks, the top left corner, which accounts for 4% of the total area is set as the target bounding box.

Evaluation metrics. We employ Intersection over Union (IoU) with a threshold of 0.5 (IoU@0.5) as the evaluation metric. A prediction is considered correct if the IoU between the predicted and ground-truth bounding boxes is greater than 0.5. For untargeted adversarial attacks, a lower IoU@0.5 value indicates a more effective attack. Conversely, for the two proposed targeted adversarial attacks, a higher IoU@0.5 value signifies a more effective attack.

3.2 Main Results

Table 1 presents the results of the two proposed untargeted adversarial attack methods, with the results without attacks serving as a baseline for comparison. Image embedding attacks reduce the average IoU@0.5 value to 23.24%, while textual bounding box attacks decrease it to an average value of 33.90%. This difference in effectiveness may be attributed to the fact that image embedding attacks disrupt the original image features, directly impacting the visual grounding capabilities of MLLMs. In contrast, textual bounding box attacks primarily affect the textual generation process of MLLMs, which might not have as significant an effect on tasks that heavily rely on visual input.

Table 2 shows the results of two proposed targeted adversarial attack paradigms. The results without attacks refer to the experiments when original images are used as inputs, with no adversarial perturbations, but with altered labels. Exclusive targeted adversarial attacks can enhance the average IoU@0.5 from 0.13% to 61.84%. Meanwhile, Permuted targeted adversarial attacks can improve the IoU@0.5 from 8.89% to 30.14%. It can be observed that permuted targeted adversarial attacks are more challenging. The reason is potentially that the area and position of target bounding box area in exclusive targeted adversarial attacks are larger and fixed, whereas the area and position of the target bounding box in permuted targeted adversarial attacks are more refined and random.

4 RELATED WORK

4.1 MULTIMODAL LARGE LANGUAGE MODELS

Multimodal large language models (MLLMs) integrate vision modalities into large language models (LLMs) to extend their capabilities, broadening their scope beyond standard textual understanding and improving their performance across various multimodal tasks (Li et al., 2022a; Zhu et al., 2023; Chen et al., 2023a; Ma et al., 2022a;b; 2024). Recent studies unlock visual grounding capabilities of MLLMs to address localization tasks with region-aware functionalities. Specifically, KOSMOS-2 (Peng et al., 2023) and VisionLLM (Wang et al., 2024a) introduce additional location tokens to the vocabulary, enabling the conversion of coordinates into textual representations, thereby enhancing regional comprehension. Moreover, Shikra (Chen et al., 2023b) and MiniGPT-v2 (Chen et al., 2023a) directly represent spatial coordinates using natural language, simplifying the integration of spatial data into the model. Despite the effective performance, the security threat for visual grounding of MLLMs, including adversarial learning (Goodfellow et al., 2015; Carlini et al., 2019; Dong et al., 2023), backdoor learning (Li et al., 2022b; Gao et al., 2023b;a; Bai et al., 2023a), poisoning learning (Shafahi et al., 2018), and Trojan learning (Rakin et al., 2020; Bai et al., 2022a; 2023b), has not been studied well.

4.2 ADVERSARIAL ATTACKS

Adversarial attacks (Goodfellow et al., 2015; Dong et al., 2018; Ilyas et al., 2018; Zhang et al., 2019; Bai et al., 2020b;a; 2021; 2022b) have been widely studied for classification models, where imperceptible and carefully crafted perturbations are applied to input data to mislead the model into producing incorrect predictions. Inspired by the adversarial vulnerability observed in vision tasks, early efforts are devoted to investigating adversarial attacks against MLLMs (Dong et al., 2023; Gao et al., 2024a; Wang et al., 2024b). However, the adversarial robustness of MLLMs with visual grounding ability is still under-explored. Since visual grounding reveals the model's perception process (Zhang et al., 2018; Li & Sigal, 2021), it can serve as a good proxy to understand the model behavior before and after the adversarial attacks. To this end, we designing effective attack methods to evaluate the adversarial robustness of MLLMs with grounding ability.

5 CONCLUSION

In this paper, we aim to craft imperceptible perturbations to generate adversarial images, evaluating the adversarial robustness for visual grounding of MLLMs. We propose three adversarial attack paradigms: untargeted adversarial attacks, exclusive targeted adversarial attacks, and permuted targeted adversarial attacks. Comprehensive experimental results on three benchmark datasets, namely RefCOCO, RefCOCO+, and RefCOCOg, demonstrate the effectiveness of our proposed attacks. We hope that our proposed adversarial attacks can serve as a baseline to evaluate the visual grounding ability in adversarial robustness of MLLMs and inspire more research to focus on visual grounding of MLLMs.

ETHICS STATEMENT

Please note that we restrict all experiments in the laboratory environment and do not support our adversarial attacks in the real scenario. The purpose of our work is to raise the awareness of the concern in availability of MLLMs and call for practitioners to pay more attention to the visual grounding in adversarial robustness of MLLMs and model trustworthy deployment.

ACKNOWLEDGEMENT

This work is supported in part by the National Natural Science Foundation of China under Grant 62171248, Shenzhen Science and Technology Program (JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1).

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In NeurIPS, 2022.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. (ab) using images and sounds for indirect instruction injection in multi-modal llms. arXiv preprint arXiv:2307.10490, 2023.

Jiawang Bai, Bin Chen, Yiming Li, Dongxian Wu, Weiwei Guo, Shu-tao Xia, and En-hui Yang. Targeted attack for deep hashing based retrieval. In ECCV, 2020a.

Jiawang Bai, Kuofeng Gao, Dihong Gong, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Hardly perceptible trojan attack against neural networks with bit flips. In ECCV, 2022a.

Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, and Shu-Tao Xia. Targeted attack against deep neural networks via flipping limited weight bits. In ICLR, 2022b.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. arXiv preprint arXiv:2311.16194, 2023a.

Jiawang Bai, Baoyuan Wu, Zhifeng Li, and Shu-Tao Xia. Versatile weight attack via flipping limited bits. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023b.

Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. Improving query efficiency of black-box adversarial attack. In ECCV, 2020b.

Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In ICLR, 2021.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, 2019.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? arXiv preprint arXiv:2306.15447, 2023.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In CVPR, 2022.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478, 2023a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023b.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In CVPR, 2018.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? arXiv preprint arXiv:2309.11751, 2023.

Kuofeng Gao, Jiawang Bai, Baoyuan Wu, Mengxi Ya, and Shu-Tao Xia. Imperceptible and robust backdoor attack in 3d point cloud. IEEE Transactions on Information Forensics and Security, 19: 1267–1282, 2023a.

Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In CVPR, 2023b.

Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In ICLR, 2024a.

Kuofeng Gao, Jindong Gu, Yang Bai, Shu-Tao Xia, Philip Torr, Wei Liu, and Zhifeng Li. Energy-latency manipulation of multi-modal large language models via verbose samples. arXiv preprint arXiv:2404.16557, 2024b.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In ICML, 2018.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In EMNLP, 2014.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In NeurIPS, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, 2022a.

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.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In NeurIPS, 2021.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023.

Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In ACM MM, 2022a.

Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. Simvtp: Simple video text pre-training with masked autoencoders. arXiv preprint arXiv:2212.03490, 2022b.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In AAAI, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In ICLR, 2018.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In CVPR, 2016.

OpenAI. Gpt-4 technical report. 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824, 2023.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak large language models. arXiv preprint arXiv:2306.13213, 2023.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In CVPR, 2020.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In NeurIPS, 2018.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In NeurIPS, 2024a.

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images. arXiv preprint arXiv:2402.14899, 2024b.

Dingcheng Yang, Yang Bai, Xiaojun Jia, Yang Liu, Xiaochun Cao, and Wenjian Yu. Cheating suffix: Targeted attack to text-to-image diffusion models with multi-modal priors. arXiv preprint arXiv:2402.01369, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In ECCV, 2016.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In CVPR, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In ICML, 2019.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. arXiv preprint arXiv:2305.16934, 2023.

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