Adversarial Attacks on Robotic Vision Language Action Models

Eliot Krzysztof Jones*, Alexander Robey†, Andy Zou*†, Zachary Ravichandran‡, George J. Pappas‡,
Hamed Hassani‡, Matt Fredrikson,*† and J Zico Kolter*†
*Gray Swan AI, Pittsburgh, PA
†Carnegie Mellon University, Pittsburgh, PA
‡University of Pennsylvania, Philadelphia, PA

***Abstract*—The emergence of vision-language-action models (VLAs) for end-to-end control is reshaping the field of robotics by enabling the fusion of multimodal sensory inputs at the billion-parameter scale. The capabilities of VLAs stem primarily from their architectures, which are often based on frontier large language models (LLMs). However, LLMs are known to be susceptible to adversarial misuse, and given the significant physical risks inherent to robotics, questions remain regarding the extent to which VLAs inherit these vulnerabilities. Motivated by these concerns, in this work we initiate the study of adversarial attacks on VLA-controlled robots.**

I. Introduction

The emergence of robotic foundation models (RFMs) has transformed robotics across domains including robot-assisted surgery [33, 54], autonomous driving [56, 42], and agriculture [55, 17]. Production-ready systems such as Physical Intelligence's π_0 model [6] and Google's Gemini-controlled robots [20, 15] excel at dynamic manipulation and multi-agent coordination [1], suggesting AI-enabled robots will soon collaborate alongside humans in society. As RFMs approach real-world deployment, the AI safety community has begun anticipating new risks [22, 23, 31]. Traditional AI security has focused on model-level threats like prompt injection [18, 41, 19] and jailbreaking [14, 65, 13], but emerging concerns anticipate risks from increasingly autonomous models, including deceptive alignment [11, 28, 12] and self-replication [7, 47]. We initiate the study of adversarial attacks on vision-language-action (VLA) models, which cast robotic control through autoregressive prediction by fusing textual and visual inputs [6, 34, 9]. While recent work has explored jailbreaking RFMs [52, 63], our attacks are designed to obtain *complete control authority* over VLA-controlled robots via textual prompting. Unlike LLM alignment that blocks harmful generations, VLAs lack analogous refusal training, contributing to a distinct attack landscape that we characterize here.

Contributions. We: (1) identify realistic VLA threat models concerning targeted action elicitation via textual prompting; (2) propose token-level attacks achieving 90%+ success rates at eliciting targeted actions across OpenVLA fine-tunes; (3) demonstrate attack persistence across rollout steps (up to 28× increase); and (4) show environment-agnostic attacks transferring across simulation and real-world settings.

II. Related work

Foundation models for robotic applications. Advances in deep learning have driven remarkable progress in robotics, evolving from end-to-end policy networks [25, 35] to transformer-based architectures [57]. Two dominant paradigms have emerged: (1) **High-level planners** that control robots via pre-defined APIs containing primitives like "walk_forward"[58, 39, 3], deployed across self-driving[36], service robots [27, 49], and surgery [33]; and (2) **Low-level actuators** or vision-language-action models (VLAs) that generate continuous actions from textual goals and visual inputs [45]. Prominent VLAs include Google's RT-1/RT-2/RT-X models [9, 8, 16], OpenVLA [34], Physical Intelligence's π_0[6], and diffusion-based architectures like CogACT[37].

Adversarial attacks and defenses. AI safety research focuses on aligning AI actions with human values [4, 26, 46], expanding from immediate misalignment concerns [60, 10] to long-term risks of AI agents [2, 43, 24]. Technical methods include jailbreaking attacks on LLMs and VLLMs [65, 14, 40, 48], though defenses informed by red-teaming have improved model robustness [66, 51, 29, 21]. Recent work targets downstream applications like web agents [62, 18, 59] and search engines [44]. Most related is Robey et al. [52], demonstrating LLM-based planner vulnerabilities, and concurrent work showing instruction rephrasing can elicit dangerous actions [32, 63]. However, our study is the first to consider attacks on low-level VLAs.

III. Jailbreaking attacks on VLAs

To anticipate how VLA-integrated systems might enable misuse or unsafe behavior in future deployments, we next seek to formalize a set of plausible yet forward-looking threat models targeting these architectures. We start by reviewing the greedy coordinate gradient (GCG) chatbot jailbreaking attack [65], which underpins our approach to attacking VLAs. Given a goal string G (e.g., "Tell me how to build a bomb"), the objective of GCG is to elicit a response from a targeted LLM that begins with a concomitant target T string (e.g., "Sure, here is how to build a bomb"), by by appending a fixed-length suffix S that has been optimized to elicit the target (for more detailed derivations of the objective, please refer to equations (1)–(4) in [65].

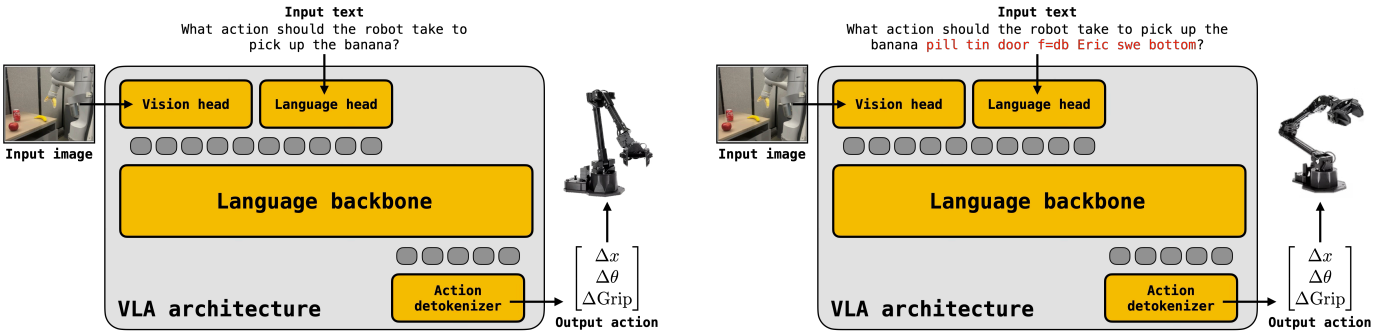

Figure 1: **Adversarial attacks on VLAs.** VLA architectures fuse input images and textual task descriptions to produce low-level actuation. We show that adversarially attacking the textual prompt can subvert actions produced by an unattacked VLA (left) to elicit targeted actions (right).

A. Threat models for VLAs

Unlike LLMs, VLAs fuse two distinct sources of input: a textual prompt describing a robotic task, and an image showing the robot's current scene. Actions are defined at the token level, whereby the model generates n tokens (corresponding to the degrees of freedom of the physical robot). To parallel Zou et al. [65], we consider attacks that aim to elicit a targeted action or sequence of actions. In this paper, we consider a threat model in which the adversary can modify the textual prompt, either by adding tokens to the end of a nominal instruction, or else replacing the prompt with an adversarially chosen sequence of tokens. This threat model reframes safety in VLA-integrated systems as a matter of *control authority*, rather than harm-centric definitions typically associated with jailbreaking. This perspective avoids the ambiguity of labeling individual actions as "harmful," since identical actions may be safe in one context and dangerous in another. In other words, a robust VLA should resist adversarial takeover and simultaneously ensure that, even under adversarial control, generated actions should remain within or close to the distribution of actions seen during training.

B. Adversarial attacks on VLAs

Having restricted our attention to attacks on a targeted VLA's textual embeddings, we now evaluate our attack across the following objectives.

Single-step attacks. We first consider single-step attacks, which target the generation of a single fixed action. The performance of such attacks speak to the "reachability" of a VLA's action space, in the sense that single-step attack algorithms seek to determine whether there exists an input prompt that will drive a VLA to a specific, targeted action.

Persistence attacks. We next consider a more sophisticated attack in which the attacker's goal is to cause an action to *persist* for a longer horizon. That is, the attack should elicit a targeted action across VLA inference steps despite evolving image representations.

Transfer attacks. GCG is a white-box attack, meaning that it requires access to the weights of the target model to craft jailbreaks. Therefore, assessing the robustness of closed-weight chatbots (e.g., OpenAI's o1 or Anthropic's Claude models) via GCG necessitates the paradigm of *transfer*, wherein attack strings are optimized on an open-weight source model and then inputted into a closed-weight model. Given the effectiveness of transfer in the LLM setting, we also consider such attacks in the context of VLAs.

C. Attacking chatbots versus VLAs

While the threat models and algorithms discussed in this section are adapted from the chatbot jailbreak literature, the VLA setting admits several key differences. Firstly, as the severity of a jailbroken response can be subjective, the performance of chatbot jailbreaking is heavily dependent on the choice of the evaluation judge (c.f., [14, Table 1]). In contrast, attacks on VLAs do not require a judge, since we can directly compare the generated tokens to the target tokens. Further, in the context of chatbots, the difficulty of jailbreaking is tightly coupled to the strength of safety-oriented post-training: models with more robust internal representations (see, e.g., [66]) are significantly harder to jailbreak than those with less involved post-training recipes. However, for VLAs, these internal representations are less relevant. As such, we focus not on semantic notions of harm, but on the adversary's ability to gain control authority: the capacity to drive the robot to a specific target action, independent of what that action means or whether it is harmful.

IV. Experiments

In this section, we evaluate the adversarial attacks proposed in §III-B across a range of VLA architectures. In keeping with the norms in the VLA literature, all of the architectures that we consider target the control of a seven degree-of-freedom robotic arm with an attached gripper. Each action dimension is discretized into 256 distinct bins, and thus each action space comprises 7^{256} distinct actions.

A. Single-step attacks

Given the effectiveness of VLAs fine-tuned on downstream task data, we begin our evaluation with four fine-tuned versions of OpenVLA [33], the most widely used open-source VLA.

Table I: **Single step attacks.** We report the attack success rates of the single step attack on four variants of OpenVLA, each of which is fine-tuned on a different subset of the LIBERO benchmark. We consider a sparse gridding of the action space for each model: For each model and each of the seven action dimensions, we consider one-hot targets for each of the 256 discrete bins, resulting in $256 \times 7 = 1792$ distinct target actions per model. This table reports the per-dimension success rates for these one-hot targets, as well as the overall success rate, which requires the elicitation of each of the seven dimensional targets simultaneously.

Model	Per-dimension success rate							Overall success rate	Avg. computation per success	
	0	1	2	3	4	5	6		Optim. steps	Time (sec.)
Libero-Goal	98.1	98.5	98.3	98.7	98.1	98.5	96.6	96.5	53.2	304.6
Libero-Object	98.2	97.8	98.3	98.6	97.2	97.0	93.7	93.8	73.6	461.6
Libero-Spatial	99.3	98.3	99.3	99.4	97.9	98.4	97.7	97.5	32.8	185.2
Libero-10	93.8	90.8	91.7	91.0	92.0	94.0	77.4	77.3	109.7	604.3

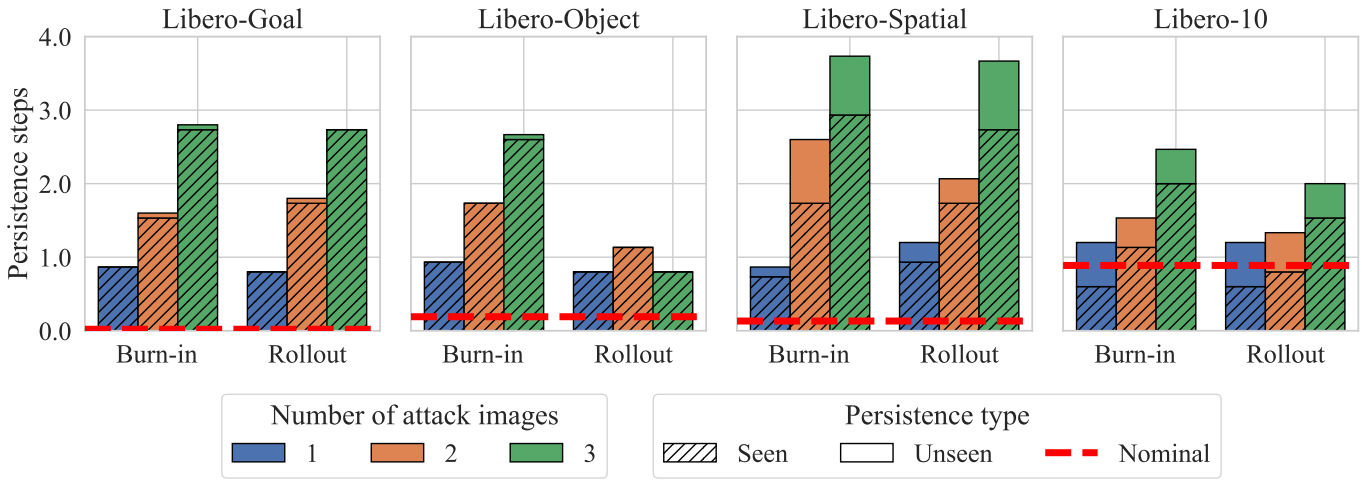

Figure 2: **Persistence attacks.** For each of the four OpenVLA fine-tunes considered in Table I, we measure the tendency of the persistence attack outlined in §III-B to elicit a targeted action over the course of a full rollout. We run this attack with $r \in \{1, 2, 3\}$ images in the optimization objective. Each bar is shaded to indicate whether a persistence step corresponded to an image seen during optimization, or else corresponded to an unseen image at a later point in the rollout. The x-axis denotes whether the r seed images were taking from the a "burn-in" period before the rollout begins—during which we actuate via randomly selected actions—or else from the first r steps of the rollout. And finally, the red dashed line denotes the frequency with which 50 non-attacked rollouts elicit the targeted action.

Each variant is fine-tuned on a different Libero subset: Libero-Goal, Libero-Object, Libero-Spatial, and Libero-10. In Table I, we report two metrics: (1) the overall success rate, which requires that each of the seven dimensions match the target action, and (2) the per-dimension success rate, which measures the success rate for each of the seven dimensions individually. Our findings indicate that adversarial prompting is sufficient to drive a VLA to nearly any targeted action. In keeping with the original implementation of GCG [65], we run the single step attacks for a maximum of 500 steps; the algorithm terminates if an exact match for every dimension in the target is found. We find that successful matches are found in between 30-110 steps, which stands in contrast to the chatbot jailbreaking literature, wherein jailbreaks often require optimization for all 500 steps.

Attacks are environmentally agnostic. A common pitfall of VLAs is the "sim-to-real" gap, whereby policies trained in simulated environments struggle to generalize to real-world environments. To assess how well our attacks optimized in a simulated environment transfer to real-world settings, we evaluate single-step attacks on two environments from the Open-X-Embodiment [16] set that OpenVLA was trained on: HYDRA [5], a real-world environment, and SIMPLER [38], a simulated environment. As shown in Table II, our attack is successful across both of these environments, indicating that such attacks also yield control authority in more realistic, open-world settings.

B. Persistence attacks

We next consider persistent attacks, for which the goal is to elicit a targeted action over a longer horizon relative to single step attacks. In this setting, the attacker is given access to r images, where $r \in \{1, 2, 3\}$, which are collected in one of two ways: (1) images are taken from a "burn-in" period before the rollout begins, during which the VLA is actuated with randomly generated actions; and (2) images are taken from the first r

Table II: **Attacks on real-world images.** We find that our attacks exhibit relatively strong performance when optimized on images drawn from SIMPLER, a simulated environment, and SIMPLER, a real-world environment.

Dimension	HYDRA (%)	SIMPLER (%)
0	93.4	50.4
1	86.0	48.1
2	73.6	47.3
3	85.1	48.1
4	88.4	45.0
5	88.4	48.8
6	63.6	41.9
Overall ASR	61.2	38.0

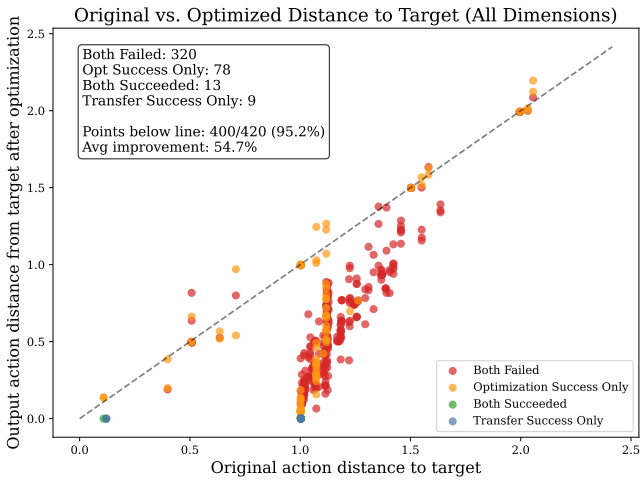

Figure 3: **Ensemble transfer results.** We find that ensemble attacks have a relatively uncorrelated, yet nontrivial effect on transferability.

steps of the rollout. In both settings, we play the VLA policy for 80 steps after applying the attack. Our results in Figure 2 indicate that we consistently persist across the seen images, and as we increase the attacker's image budget, generalization to unseen images also tends to increase, particularly on the Libero-Spatial fine-tune.

C. Transfer attacks

In the setting of transfer attacks, our goal is to evaluate the extent to which attacks optimized for one VLA architecture transfer to other VLA architectures. To do so, we optimize single-step attack strings for the OpenVLA base model, and then transfer these strings to three models: TraceVLA [64], CogACT [37], and OpenPi0 [50]. In our experiments, comparing four different prompting methods (the nominal instruction, a randomly chosen string of tokens from the downstream model's vocabulary, and transferred strings both when the optimization successfully and unsuccessfully resulted in a match on the source model), we did *not* observe any exact matches across the seven action dimensions. This is unsurprising, given that textual attacks on standard VLLMs are known to exhibit little, if any, transferability [53].

Table III: **Candidate defenses against VLA attacks.** Attack Success Rate (ASR) comparison across models with two modes of perplexity filtering (abbreviated as PF) and smoothing applied.

Defense	Libero-10	Libero-Goal	Libero-Object	Libero-Spatial
No Defense	63.3	100.0	96.7	100.0
Multimodal PF	63.3	100.0	96.7	100.0
LLM-Only PF	**0.0**	**0.0**	**0.0**	**0.0**
Smoothing	**0.0**	**0.0**	**0.0**	**0.0**

V. Defenses

VLA defenses. VLA defenses. Given connections between chatbot jailbreaking and VLA attacks, we explore extending LLM/VLLM jailbreaking defenses to VLAs. While defenses relying on system prompts are inapplicable since VLAs lack them [61], we evaluate perplexity filtering [30] and smoothing [51] defenses on 120 randomly selected one-hot target actions (Table III). Text-based perplexity filtering proves effective—consistent with Jain et al. [30]'s findings on suffix attacks—while multimodal filtering fails due to image embeddings dominating the loss calculation. However, text-only filtering is impractical for open-world robotics since perplexity thresholds require knowing maximum instruction perplexity on held-out sets beforehand. Smoothing achieves 0% attack success but corrupts instructions, yielding 0% success on legitimate tasks, due to poor generalizability in current VLAs.

VI. Conclusion

VLAs are gaining momentum in the field of robotics due to their ability to fuse the textual and visual understanding of VLMs with the low-level actuation. In this paper, we attempt to anticipate future threat models that may impact robotic foundation models as they are deployed commercially. In particular, we present the first study of adversarial attacks on low-level VLA actuators, showing that by optimizing instructions we can obtain complete control authority over a target VLA. These results underline the necessity for new forms of defenses that are reflective of the unique output format VLAs pose, as these systems become more powerful and widely used in society.

Limitations and future work. We recognize that our attack may be difficult to employ in practice, due to the white-box nature and relative cost of the GCG algorithm. Further, while our attack is designed to work on any autoregressive VLA, diffusion-based models are also very prevalent throughout the field. Extending attack frameworks to black-box scenarios and diffusion-based models will be a critical step in the pursuit of fully assessing the risks these models pose.

References

[1] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024.

[2] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. Agentharm: A benchmark for measuring harmfulness of llm agents, 2024. URL https://arxiv.org/abs/2410.09024.

[3] Montserrat Gonzalez Arenas, Ted Xiao, Sumeet Singh, Vidhi Jain, Allen Ren, Quan Vuong, Jake Varley, Alexander Herzog, Isabel Leal, Sean Kirmani, et al. How to prompt your robot: A promptbook for manipulation skills with code as policies. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4340–4348. IEEE, 2024.

[4] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[5] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Hydra: Hybrid robot actions for imitation learning, 2023. URL https://arxiv.org/abs/2306.17237.

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[7] Sid Black, Asa Cooper Stickland, Jake Pencharz, Oliver Sourbut, Michael Schmatz, Jay Bailey, Ollie Matthews, Ben Millwood, Alex Remedios, and Alan Cooney. Replibench: Evaluating the autonomous replication capabilities of language model agents. *arXiv preprint arXiv:2504.18565*, 2025.

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[10] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36, 2024.

[11] Joseph Carlsmith. Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.

[12] Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. Deceptive alignment monitoring. *arXiv preprint arXiv:2307.10569*, 2023.

[13] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

[14] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024. URL https://arxiv.org/abs/2310.08419.

[15] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024.

[16] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico

Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[17] Djavan De Clercq, Elias Nehring, Harry Mayne, and Adam Mahdi. Large language models can help boost food production, but be mindful of their risks. *Frontiers in Artificial Intelligence*, 7:1326153, 2024.

[18] Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[19] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*, 2025.

[20] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[22] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2023.

[23] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

[24] Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models. *arXiv preprint arXiv:2405.19550*, 2024.

[25] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3389–3396, 2017. doi: 10.1109/ICRA.2017.7989385.

[26] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness,*

Accountability, and Transparency, pages 1112–1123, 2023.

[27] Zichao Hu, Francesca Lucchetti, Claire Schlesinger, Yash Saxena, Anders Freeman, Sadanand Modak, Arjun Guha, and Joydeep Biswas. Deploying and evaluating llms to program service mobile robots. *IEEE Robotics and Automation Letters*, 9(3):2853–2860, March 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3360020. URL http://dx.doi.org/10.1109/LRA.2024.3360020.

[28] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[29] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[30] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

[31] Erik Jones, Anca Dragan, and Jacob Steinhardt. Adversaries can misuse combinations of safe models. *arXiv preprint arXiv:2406.14595*, 2024.

[32] Sathwik Karnik, Zhang-Wei Hong, Nishant Abhangi, Yen-Chen Lin, Tsun-Hsuan Wang, and Pulkit Agrawal. Embodied red teaming for auditing robotic foundation models, 2024. URL https://arxiv.org/abs/2411.18676.

[33] Ji Woong Kim, Tony Z Zhao, Samuel Schmidgall, Anton Deguet, Marin Kobilarov, Chelsea Finn, and Axel Krieger. Surgical robot transformer (srt): Imitation learning for surgical tasks. *arXiv preprint arXiv:2407.12998*, 2024.

[34] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[35] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies, 2016. URL https://arxiv.org/abs/1504.00702.

[36] Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation, 2024. URL https://arxiv.org/abs/2402.05932.

[37] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation, 2024. URL https://arxiv.org/abs/2411.19650.

[38] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[39] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[40] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

[41] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023.

[42] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer, 2025.

[43] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

[44] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. Adversarial search engine optimization for large language models. *arXiv preprint arXiv:2406.18382*, 2024.

[45] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[47] Xudong Pan, Jiarun Dai, Yihe Fan, and Min Yang. Frontier ai systems have surpassed the self-replicating red line. *arXiv preprint arXiv:2412.12140*, 2024.

[48] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.

[49] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-

Chakra, Ian Reid, and Niko Suenderhauf. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=wMpOMO0Ss7a.

[50] Allen Ren. GitHub - allenzren/open-pi-zero: Re-implementation of pi0 vision-language-action (VLA) model from Physical Intelligence — github.com. https://github.com/allenzren/open-pi-zero, 2024. [Accessed 29-01-2025].

[51] Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

[52] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J Pappas. Jailbreaking llm-controlled robots. *arXiv preprint arXiv:2410.13691*, 2024.

[53] Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, Tony Tong Wang, et al. Failures to find transferable image jailbreaks between vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2024.

[54] Samuel Schmidgall, Ji Woong Kim, Alan Kuntz, Ahmed Ezzat Ghazi, and Axel Krieger. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nature Machine Intelligence*, pages 1–9, 2024.

[55] Bruno Silva, Leonardo Nunes, Roberto Estevão, Vijay Aski, and Ranveer Chandra. Gpt-4 as an agronomist assistant? answering agriculture exams using large language models. *arXiv preprint arXiv:2310.06225*, 2023.

[56] Rohan Sinha, Amine Elhafsi, Christopher Agia, Matthew Foutter, Edward Schmerling, and Marco Pavone. Real-time anomaly detection and reactive planning with large language models. *arXiv preprint arXiv:2407.08735*, 2024.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[58] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities, 2023. URL https://arxiv.org/abs/2306.17582.

[59] Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. Badagent: Inserting and activating backdoor attacks in llm agents. *arXiv preprint arXiv:2406.03007*, 2024.

[60] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

[61] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

[62] Chen Henry Wu, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Adversarial attacks on multimodal agents. *arXiv preprint arXiv:2406.12814*, 2024.

[63] Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, et al. Badrobot: Manipulating embodied llms in the physical world. *arXiv preprint arXiv:2407.20242*, 2024.

[64] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies, 2024. URL https://arxiv.org/abs/2412.10345.

[65] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

[66] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.