Future Directions for Quality Diversity Red-Teaming of Vision-Language-Action Models

Siddharth Srikanth[12], Freddie Liang[1], Ya-Chuan Hsu[1], Varun Bhatt[1], Shihan Zhao[1], Henry Chen[1], Bryon Tjanaka[1], Minjune Hwang[1], Akanksha Saran[3], Daniel Seita[1†], Aaquib Tabrez[4†], Stefanos Nikolaidis[1†]
[1]University of Southern California [2]Columbia University [3]Adobe Research
[4]Cornell University [†]Equal advisement

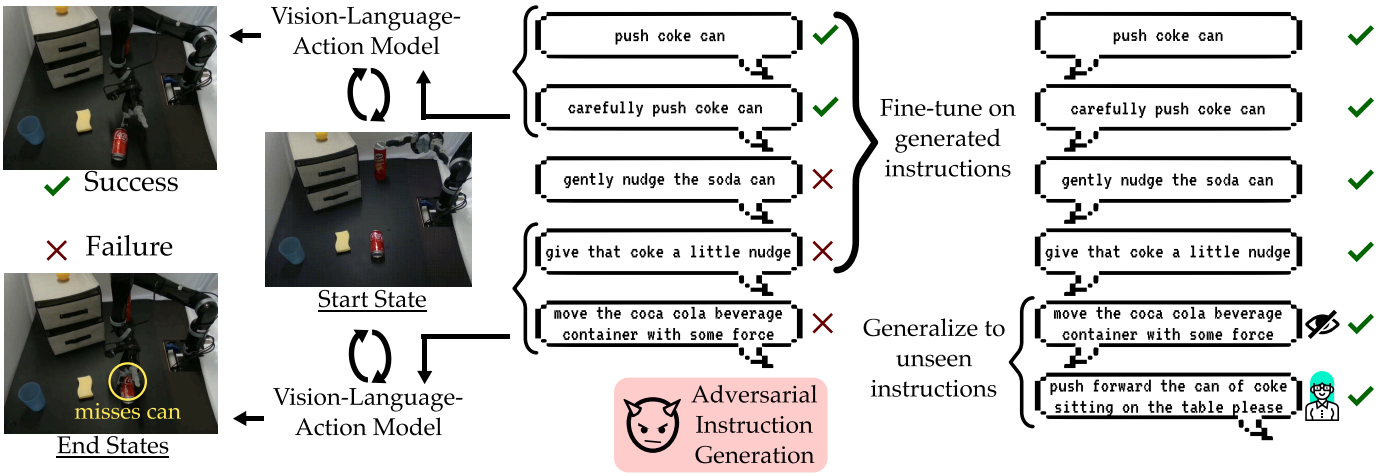


Fig. 1: Our framework, Q-DIG, aims to make VLA-powered robots robust to different instruction wordings by generating adversarial instructions that induce task failures. For example, a VLA might succeed when instructed to "push coke can" (the VLA uses one of the end effectors to push the can), but fail when instructed "meticulously exert force upon the aluminum beverage container" (the VLA misses the coke can and pushes nothing). Augmenting demonstrations with such diverse instructions and fine-tuning the VLA using them leads to better generalization to unseen and human-generated instructions.

***Abstract*—Vision-Language-Action (VLA) models have significant potential to enable general-purpose robotic systems for a range of vision-language tasks. However, the performance of VLA-based robots is highly sensitive to the precise wording of language instructions, and it remains difficult to predict when such robots will fail. In our prior work, we introduced Q-DIG, a novel framework that combines Quality Diversity (QD) optimization with Vision-Language Models (VLMs) to generate a broad spectrum of adversarial instructions that expose meaningful vulnerabilities in VLA behavior. We evaluated Q-DIG across three state-of-the-art VLA models, conducted a user study to show the human-likeness of Q-DIG instructions over baselines, and validated our framework with a sim-to-real evaluation in the real-world. In this work, we summarize the Q-DIG framework and key findings, and propose three new applications of its work: (1) Searching for Environmental Diversity, (2) Surrogate Modeling for Efficient Evaluation, and (3) Red-Teaming Hierarchical VLA Models.**

I. Introduction and Motivation

Vision-Language-Action (VLA) models, which fine-tune Vision-Language Models (VLMs) to produce robot actions, have demonstrated impressive generalization in robotics tasks [10, 13, 42], such as out-of-the-box deployment across novel environments and efficient learning of new tasks with limited data [1, 17, 5, 14].

Despite this progress, VLAs remain vulnerable to jailbreaking [33] and red-teaming [27] attacks through adversarial instructions, which can elicit unexpected failures [40] and limit deployment in safety-critical applications. In part, this arises because current VLA datasets and fine-tuning protocols provide only a narrow mapping between language instructions and demonstrations. For example, LIBERO [21] provides a single instruction per task with multiple demonstrations, DROID [16] associates each trajectory with at most three instructions, and Bridge [39] has a one-to-one mapping between instructions and demonstrations. For example, a VLA might successfully complete the task of pushing a Coke can when the instruction is exactly "push Coke can," while the instruction "gently nudge the soda can!" causes the VLA to fail.

We study this problem through the lens of *red-teaming*: the systematic discovery of inputs that expose failures in a model [27, 12, 31]. Prior work [33] showed that jailbroken LLM-controlled robots can be induced to take harmful actions across self-driving, unmanned vehicle navigation, and quadruped control. While prior methods have been developed to induce failures in VLAs via adversarial instructions, it is still challenging to generate *realistic* instructions that elicit a *controllably diverse* range of failures. In particular, Embodied Red Teaming (ERT) [15] leverages in-context learning [7] to generate adversarial instructions, but does not explicitly target a set of failure modes specified by system designers. Additionally, some generated instructions may fall outside the

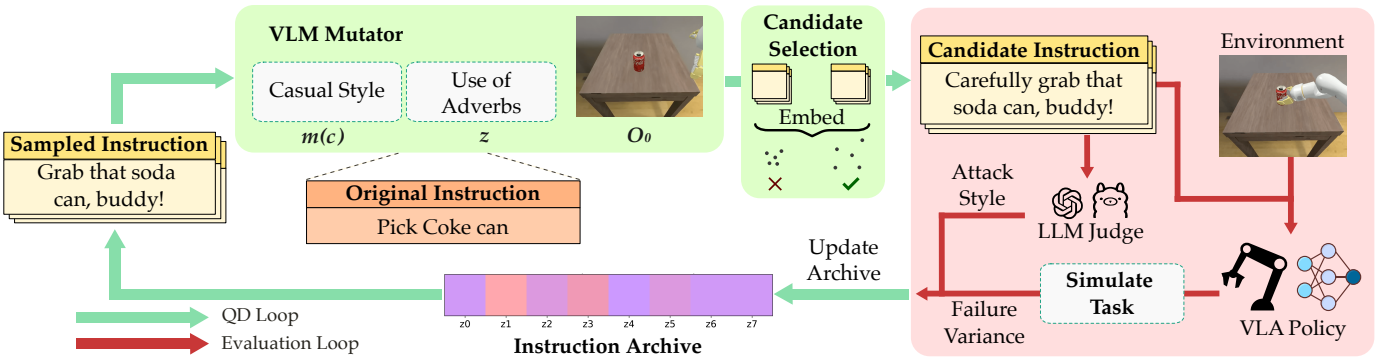


Fig. 2: Overview of Q-DIG. Q-DIG leverages previously generated instructions as in-context examples to generate new adversarial instructions in target attack styles (green arrows). The generated instructions are evaluated (red arrows) to obtain the variance of failure rates they induce in the VLA as well as their actual attack style. Instructions inducing high failure rates with different attack styles (**z0** to **z7** from Table V) are stored in an archive, providing high-quality and diverse examples for future iterations.

distribution of instructions that users normally provide. For the Coke can task above, one such unrealistic instruction would be "spot the red and white can lying flat and push it with precision."

We argue that Quality Diversity (QD) optimization [30, 29, 23] provides a natural framework for red teaming embodied models by leveraging previously found solutions as stepping stones to generate new solutions that are higher performing or more diverse. The QD works closest to ours are Rainbow Teaming [34] and PLAN-QD [36]. Rainbow Teaming uses LLMs as mutators to generate prompt variants according to semantic failure modes, and PLAN-QD searches for diverse prompts that induce varied collaborative behaviors in LLM-powered agents for sequential decision-making tasks. While both Rainbow Teaming and PLAN-QD advance systematic prompt search, they lack grounding in physical tasks and visual information required for embodied domains.

Quality Diversity (QD) algorithms search for a diverse collection of high-performing solutions [30] by leveraging previously found solutions as stepping stones to generate new solutions that are higher performing or more diverse. QD has been applied in robotics, including training diverse RL agents [28, 24, 38, 2]. QD has also been applied to generate diverse scenarios for autonomous agents [3] and deployed robotic systems [11, 4]. More recently, previous work has applied QD to text generation with LLMs [22, 9, 6, 20], where previously discovered prompts [9] or poems [6] serve as context for an LLM to generate new and improved variants,

In our prior work [37], we introduced Q-DIG (**Q**uality Diversity for **D**iverse **I**nstruction **G**eneration) (Figure 1), a novel framework for red-teaming and improving VLAs. *Our key insight is that QD optimization combined with vision-language grounding enables a systematic discovery of diverse, realistic, and task-relevant failure modes in embodied systems.*

In this paper, we summarize the following contributions: **(1)** We present Q-DIG, a framework that leverages Quality Diversity optimization for diverse and in-distribution adversarial instruction generation. **(2)** We evaluate Q-DIG in two simulation domains, SimplerEnv [18] and LIBERO [21], and show that compared to prior VLA red-teaming methods, Q-DIG produces more diverse and in-distribution adversarial instructions. **(3)** Through a user study, we show that instructions generated by Q-DIG are more human-like compared to those generated by prior methods. **(4)** We demonstrate that fine-tuning with the augmented dataset containing adversarial instructions improves robustness to unseen instructions. **(5)** We validate our approach in a sim-to-real setting, showing that the benefits of adversarial fine-tuning transfer to real-world robotic deployments. We then propose three new applications of its work: (1) Searching for Environmental Diversity, (2) Surrogate Modeling for Efficient Evaluation, and (3) Red-Teaming Hierarchical VLA Models.

II. Summary of the Q-DIG Framework

The central idea of Q-DIG is to use QD to automatically generate a wide range of semantically diverse instructions that, when executed on the base VLA, serve as adversarial prompts revealing failure modes. In this section, we first describe how we formulate generating adversarial instructions for VLAs as a QD problem, then provide the detailed framework of QD optimization, and finally explain how the resulting instructions can be leveraged for fine-tuning. Figure 2 presents an overview of our pipeline.

A. Generating Adversarial Instructions

To discover diverse adversarial instructions for VLAs that cause failures, Q-DIG algorithmically searches for them using QD optimization. The optimization begins with the original task instruction $\boldsymbol{c}_{0,T}$, such as "Pick Coke can" in Figure 2. As the archive fills, Q-DIG samples a filled cell and retrieves its stored instruction, which then serves as a "stepping stone" for generating new adversarial instructions.

The mutator VLM is queried with an existing adversarial instruction $\boldsymbol{c}$, its attack style $m(\boldsymbol{c})$, initial observation for the task $\boldsymbol{o}_0$ (visual), and a target attack style category z to generate a candidate instruction in the target attack style (e.g., "use of adverbs" in Fig. 2). We generate k sets of candidate instructions of batch size b each, and select the one with the the most diverse instructions (i.e. the least pairwise Sentence-BERT [32] similarity).

TABLE I: Fine-tuned VLA results on the LIBERO-Goal task suite, using (i) original and (ii) adversarial plus original instructions across three VLAs: OpenVLA-OFT, $\pi_{0.5}$, and GR00T N1.6. "Original" represents evaluation on the original task instruction. Each instruction was evaluated 50 times. "Rephrase," "ERT," and "Q-DIG" represent evaluation on 2 unseen adversarial instructions. Columns correspond to the original VLA and models fine-tuned with instructions from the corresponding algorithm, and rows correspond to unseen instructions during training. Models trained on adversarial instructions (ERT and Q-DIG rows) perform better on unseen adversarial instructions.

Unseen Prompts	**OpenVLA-OFT**				$\pi_{0.5}$				**GR00T N1.6**			
	Orig	Reph	ERT	Q-DIG	Orig	Reph	ERT	Q-DIG	Orig	Reph	ERT	Q-DIG
Original	**97.4**	62.2	94.6	93.8	**96.8**	**96.8**	96.4	96.4	**77.0**	51.6	46.6	64.8
Rephrase	90.3	63.4	**90.8**	84.7	90.7	**96.0**	95.9	95.7	28.2	44.4	30.1	**47.7**
ERT	66.2	53.3	**91.1**	66.9	62.3	**72.7**	69.6	70.8	18.1	36.1	**41.9**	37.6
Q-DIG	63.9	47.6	76.4	**88.9**	67.5	71.9	**74.8**	73.4	33.7	42.4	33.6	**55.1**
Avg	76.9	55.8	**87.3**	82.1	76.8	**82.6**	82.4	82.3	33.9	42.5	36.8	**49.4**

Q-DIG evaluates this instruction by rolling out the base VLA $\pi(\boldsymbol{o}, \boldsymbol{c})$ on task T to compute the failure variance $J(\boldsymbol{c})$. In addition, as our *measure function* m, an external LLM (referred to as *LLM judge*) categorizes the instruction into one of the semantic attack styles $\{z_i\}_{i=0}^{M}$. The instruction then replaces a cell in the archive with the same style if the cell is empty or if it achieves a higher objective value than the one currently stored.

Through this iterative process of ***instruction selection***, ***mutation***, ***evaluation***, and ***archive update***, Q-DIG fills the archive with semantically diverse attack styles and failure-inducing task instructions.

B. Fine-Tuning the VLA

Once the set of adversarial instructions is discovered, we fine-tune the base VLA to mitigate the induced failures and improve the robot's robustness to these attack styles. We assume access to a dataset of expert demonstrations $\mathcal{D} = \{\tau_1, \ldots, \tau_K\}$ of size K, with each demonstration mapping to an original language instruction for the given task T. We then augment this dataset by associating the language instructions for some demonstrations with one of our generated adversarial instructions, *without* collecting additional expert demonstrations. In our experiments, we have $N = 50$ demonstrations, and $G = 9$ unique instructions (8 adversarial instructions + 1 original instruction), for each task. The augmented dataset is agnostic to the fine-tuning method. Once finetuned on this dataset, the VLA becomes more robust to different task descriptions.

III. Results and Key Takeaways

We discuss our results for the prompt generation and user studies (Sec. III-A), fine-tuning three VLAs (Sec. III-B), and real-world deployment (Sec. III-C).

A. Results: Prompt Generation

Compared to Rephrase and ERT, two standard baselines, Q-DIG generated more diverse instructions and achieved a wider coverage of failure categories across both benchmarks (Fig. 3, Table II). While all methods achieved similar BLEU diversity, Q-DIG more effectively targeted a wider range of realistic failure modes, demonstrating the benefits of quality-diversity search for adversarial instruction generation.

We conducted a user study ($n = 40$) to measure whether Q-DIG generated instructions were human-like. First, participants were given initial and goal-state images of each of the five SimplerEnv tasks, and asked to provide a normal and confusing instruction for the robot to complete the task. Then, participants were shown three variations of the initial task instruction (one from each baseline in a randomized order), and were asked to (1) rank them in decreasing order of human-likeness, and (2) rate the human-likeness of each instruction on a 7-point Likert scale. A Kruskal-Wallis test followed by post-hoc Dunn's test with Bonferroni correction showed Q-DIG's ranking to be significantly better ($p < 0.001$) than baselines. Furthermore, Likert scores indicated Q-DIG-generated instructions were significantly more human-like than ERT-generated instructions (mean difference of 1.07; $p < 0.001$).

TABLE II: Archive Coverage of failure modes (Table V) and pairwise BLEU diversity results across all 3 methods and 2 benchmarks for our OpenVLA-OFT fine-tuned model. We show the values along with their standard errors in subscripts.

Metric (Domain)	Rephrase	ERT	Q-DIG
BLEU (LIBERO)	$\mathbf{0.963}_{0.003}$	$0.951_{0.01}$	$0.947_{0.01}$
BLEU (SimplerEnv)	$0.928_{0.01}$	$0.936_{0.01}$	$\mathbf{0.946}_{0.01}$
Coverage (LIBERO)	$0.363_{0.02}$	$0.325_{0.01}$	$\mathbf{0.972}_{0.02}$
Coverage (SimplerEnv)	$0.388_{0.03}$	$0.325_{0.02}$	$\mathbf{0.913}_{0.02}$

B. Results: VLA Fine-Tuning

Across OpenVLA [17], $\pi_{0.5}$ [14], and GR00T N1.6 [25], we augment fine-tuning datasets with adversarial instructions from all three methods. See the full paper [37] for details on the evaluation. We found data augmentation improved performance on unseen adversarial prompts compared to the base models, with improvements ranging from 5–15% on average. We further observed similar robustness improvements in SimplerEnv tasks, suggesting that adversarial instruction augmentation can improve generalization beyond the original evaluation setting. These results indicate that augmenting training data with diverse adversarial instructions is an effective strategy for improving VLA robustness.

C. Results: Real-World Experiments

We evaluated several adversarial instructions generated by Q-DIG on a Gen-2 Kinova JACO arm and two RealSense Depth Camera D435i (third person and wrist view). We chose two representative tasks from our simulation experiments, "push the coke can" and "push the sponge", and ran three Q-DIG iterations to produce an archive of adversarial task

instructions for evaluation. See Figure 1 for a visual of our real-world setup. We observed that instructions that induced failures in simulation also reduced success rates in the real world, while non-adversarial instructions remained successful (Table III). We further fine-tuned the policy using Q-DIG-generated data and found improved robustness to previously unseen adversarial instructions in real-world evaluation (Table IV), suggesting that adversarial instructions discovered in simulation can be used to improve real-world VLA robustness. See the full paper [37] for detailed results.

TABLE III: Comparison of Original and Q-DIG instruction variations for the "push the coke can" real-world task.

	Original	P1	P2	P3	P4
Real-World	9/10	0/10	8/10	5/10	10/10
Simulation	10/10	4/10	7/10	4/10	10/10

TABLE IV: Unseen prompt evaluated on Original vs. Original and Q-DIG augmented models for "push the coke can" real-world task.

	Unseen P1	Unseen P2
Original	0/10	8/10
Q-DIG	7/10	9/10

IV. Future Applications of Q-DIG and Conclusion

A1. Searching for Environmental Diversity. Instead of limiting to instruction diversity, Q-DIG could be extended to explore environmental diversity. While our current framework searches over natural language variations, many VLA failures arise from changes in the visual scene itself, such as object placements and distractor objects. Benchmarks such as LIBERO-Plus [8] and VLABench [41] provide valuable testbeds for measuring generalization across tasks and scenes. Furthermore, work on unsupervised environment design (UED) automatically generates environments [26] but has not yet been applied to VLA policies. Future versions of Q-DIG could jointly optimize over both language *and* environmental perturbations to discover multimodal failure modes that are difficult to identify through instruction generation alone.

For example, a task that succeeds under a clean tabletop configuration may fail when additional visually similar objects are introduced or when target objects are partially occluded. By incorporating environmental parameters into the QD search process, Q-DIG could discover a broader range of realistic failure scenarios beyond instruction-level variation and provide more comprehensive evaluations of VLA robustness.

A2. Surrogate Modeling for Efficient Evaluation. One limitation of Q-DIG is that it evaluates the success rate induced for a given instruction by rolling out the VLA in the environment multiple times. Since VLA rollouts are computationally expensive, we were only able to run Q-DIG for 3–12 iterations.

To remedy this, prior work [3, 4] explores using surrogate modeling to predict human and robot behaviors for scenario generation systems. Future work for Q-DIG could explore training a surrogate model that predicts the failure rate of generated instructions based on the trajectories of previously evaluated examples. Rather than executing every candidate instruction in the environment, the surrogate model could estimate failure likelihood and uncertainty based on the given task and other information about the instruction, mitigating expensive VLA rollouts.

This approach could significantly improve the scalability of Q-DIG by enabling longer optimization runs and exploration of larger search spaces. Currently, Q-DIG operations in a total space of eight failure modes, whereas surrogate modeling gives promise to increase this number. Surrogate-assisted QD algorithms have been shown to improve sample efficiency in other domains, suggesting a promising direction for scaling adversarial instruction generation to larger VLA models, more complex environments, and longer-horizon tasks.

A3. Red-Teaming Hierarchical VLA Models. Recent work has explored hierarchical VLA and VLM-based planning systems that decompose high-level goals into intermediate subgoals and executable actions [35, 19]. For example, the task "make a sandwich" may first be decomposed into "pick up a slice of bread," followed by "place the bread on the counter," and then subsequent manipulation subtasks. Such hierarchical decomposition enables improved performance on long-horizon tasks by simplifying the decision-making process for the underlying VLA.

While promising, hierarchical systems introduce new failure modes: errors in early subtasks may propagate through the execution chain, leading to compounding mistakes or incorrect plans. Furthermore, ambiguity in high-level instructions may result in flawed decompositions that are difficult to recover from. For example, the order of picking between bread, lettuce, and meat for the "make a sandwich" may be ambiguous to a VLA, but more clear to humans. In this setting, Q-DIG could be used to generate instructions that specifically target weaknesses in task decomposition and hierarchical planning.

A. Conclusion

While VLA models have demonstrated remarkable progress toward general-purpose robotic systems, their sensitivity to natural language instructions inhibits them from reliable deployment. In our prior work, we introduced Q-DIG, a framework leveraging QD for foundation models to generate diverse and human-like adversarial instructions that expose vulnerabilities in VLA systems. Our results showed that, compared to baselines, Q-DIG generates visually-grounded, more diverse, more human-like adversarial instructions while remaining in-distribution to expose meaningful vulnerabilities in VLAs. We also demonstrate that augmenting fine-tuning with adversarial prompts improves VLA robustness to unseen instructions.

Looking forward, we believe QD search provides a promising foundation for scalable red-teaming of embodied AI systems. Extending Q-DIG to search over environmental variations, incorporating surrogate models for efficient evaluation, and adapting the framework to hierarchical VLA architectures represent one of several promising directions for future research. We hope that our work paves the way for future research on scalable red-teaming methods and more reliable and trustworthy robots.

ACKNOWLEDGMENTS

This work was supported by NSF CAREER #2145077 and DARPA EMHAT HR00112490409.

REFERENCES

[1] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

[2] Sumeet Batra, Bryon Tjanaka, Matthew Christopher Fontaine, Aleksei Petrenko, Stefanos Nikolaidis, and Gaurav S. Sukhatme. Proximal policy gradient arborescence for quality diversity reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[3] Varun Bhatt, Bryon Tjanaka, Matthew Fontaine, and Stefanos Nikolaidis. Deep surrogate assisted generation of environments. *Advances in Neural Information Processing Systems*, 35:37762–37777, 2022.

[4] Varun Bhatt, Heramb Nemlekar, Matthew C. Fontaine, Bryon Tjanaka, Hejia Zhang, Ya-Chuan Hsu, and Stefanos Nikolaidis. Surrogate Assisted Generation of Human-Robot Interaction Scenarios. In *Conference on Robot Learning (CoRL)*, 2023.

[5] 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. In *Robotics: Science and Systems (RSS)*, 2025.

[6] Herbie Bradley, Andrew Dai, Hannah Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Grégory Schott, and Joel Lehman. Quality-Diversity Through AI Feedback. In *International Conference on Learning Representations (ICLR)*, 2024.

[7] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[8] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. *arXiv preprint arXiv:2510.13626*, 2025.

[9] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

[10] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. Foundation Models in Robotics: Applications, Challenges, and the Future. *arXiv preprint arXiv:2312.07843*, 2023.

[11] Matthew C. Fontaine and Stefanos Nikolaidis. A quality diversity approach to automatically generating human-robot interaction scenarios in shared autonomy. In *Robotics: Science and Systems (RSS)*, 2021.

[12] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[13] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Shibo Zhao, Yu Quan Chong, Chen Wang, Katia Sycara, Matthew Johnson-Roberson, Dhruv Batra, Xiaolong Wang, Sebastian Scherer, Zsolt Kira, Fei Xia, and Yonatan Bisk. Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis. *arXiv preprint arXiv:2312.08782*, 2023.

[14] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *Conference on Robot Learning (CoRL)*, 2025.

[15] Sathwik Karnik, Zhang-Wei Hong, Nishant Abhangi, Yen-Chen Lin, Tsun-Hsuan Wang, and Pulkit Agrawal. Embodied Red Teaming for Auditing Robotic Foundation Models. *arXiv preprint arXiv:2411.18676*, 2024.

[16] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao,

Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Robotics: Science and Systems (RSS)*, 2024.

[17] 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. In *Conference on Robot Learning (CoRL)*, 2024.

[18] 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. In *Conference on Robot Learning (CoRL)*, 2024.

[19] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, and Ankit Goyal. Hamster: Hierarchical action models for open-world robot manipulation, 2025. URL https://arxiv.org/abs/2502.05485.

[20] Bryan Lim, Manon Flageat, and Antoine Cully. Large language models as in-context ai generators for quality-diversity. *arXiv preprint arXiv:2404.15794*, 2024.

[21] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Neural Information Processing Systems*, 2023.

[22] Elliot Meyerson, Mark J Nelson, Herbie Bradley, Adam Gaier, Arash Moradi, Amy K Hoover, and Joel Lehman. Language model crossover: Variation through few-shot prompting. *arXiv preprint arXiv:2302.12170*, 2023.

[23] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

[24] Olle Nilsson and Antoine Cully. Policy gradient assisted map-elites. In *Conference on Genetic and Evolutionary Computation*, 2021.

[25] NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[26] Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning (ICML)*, 2022.

[27] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[28] Thomas Pierrot, Valentin Macé, Felix Chalumeau, Arthur Flajolet, Geoffrey Cideron, Karim Beguir, Antoine Cully, Olivier Sigaud, and Nicolas Perrin-Gilbert. Diversity policy gradient for sample efficient quality-diversity optimization. In *Conference on Genetic and Evolutionary Computation*, 2022.

[29] Justin K Pugh, Lisa B Soros, Paul A Szerlip, and Kenneth O Stanley. Confronting the challenge of quality diversity. In *Conference on Genetic and Evolutionary Computation*, 2015.

[30] Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 2016.

[31] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[33] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas. Jailbreaking LLM-Controlled Robots. In *International Conference on Robotics and Automation (ICRA)*, 2025.

[34] Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Fo-

erster, Tim Rocktäschel, and Roberta Raileanu. Rainbow Teaming: Open-Ended Generation of Diverse Adversarial Prompts. In *Neural Information Processing Systems*, 2024.

[35] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[36] Siddharth Srikanth, Varun Bhatt, Boshen Zhang, Werner Hager, Charles Michael Lewis, Katia P Sycara, Aaquib Tabrez, and Stefanos Nikolaidis. Algorithmic prompt generation for diverse human-like teaming and communication with large language models. *arXiv preprint arXiv:2504.03991*, 2025.

[37] Siddharth Srikanth, Freddie Liang, Ya-Chuan Hsu, Varun Bhatt, Shihan Zhao, Henry Chen, Bryon Tjanaka, Minjune Hwang, Akanksha Saran, Daniel Seita, Aaquib Tabrez, and Stefanos Nikolaidis. Red-teaming vision-language-action models via quality diversity prompt generation for robust robot policies, 2026. URL https://arxiv.org/abs/2603.12510.

[38] Bryon Tjanaka, Matthew C. Fontaine, Julian Togelius, and Stefanos Nikolaidis. Approximating gradients for differentiable quality diversity in reinforcement learning. In *Conference on Genetic and Evolutionary Computation*, 2022.

[39] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A Dataset for Robot Learning at Scale. In *Conference on Robot Learning (CoRL)*, 2023.

[40] J. Wang, M. Leonard, K. Daniilidis, D. Jayaraman, and E. S. Hu. Evaluating pi0 in the Wild: Strengths, Problems, and the Future of Generalist Robot Policies, 2025.

[41] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. VLABench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[42] Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, Yuanpei Chen, and Yaodong Yang. A Survey on Vision-Language-Action Models: An Action Tokenization Perspective. *arXiv preprint arXiv:2507.01925*, 2025.

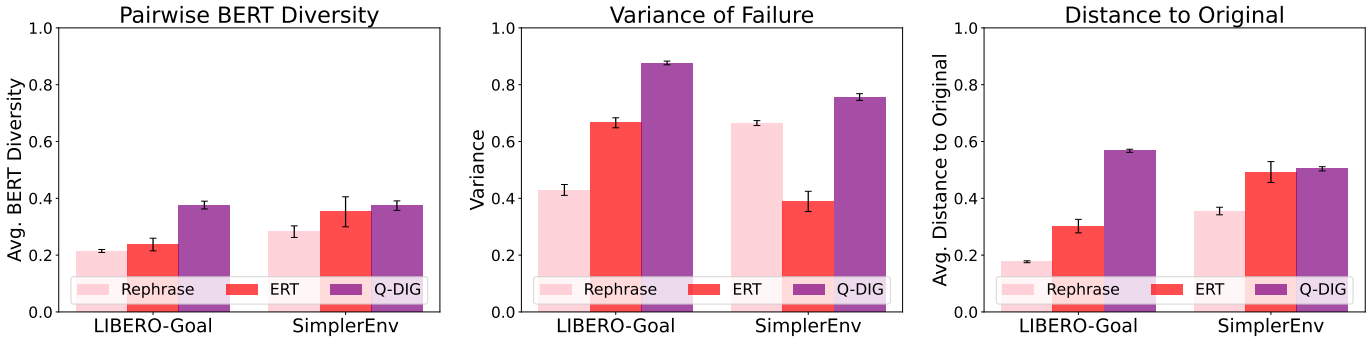
APPENDIX

A. Problem Definition

To improve the robustness of VLAs to language instructions, we address the problem of generating diverse instructions for a given task. We assume a given *base VLA*, $\pi(\boldsymbol{o}, \boldsymbol{c})$, that maps observations $\boldsymbol{o}$ (visual and/or proprioceptive) and language instructions $\boldsymbol{c} \in \mathcal{C}$ (where $\mathcal{C}$ is the set of all natural language instructions) to actions that a robot can execute. The language instructions describe a task $T \equiv (\mu_0, g_T)$, with μ_0 being the initial state distribution and $g_T(\zeta)$ being the goal predicate that outputs 1 when a rollout trajectory ζ completes the task. We define a set of language instructions $\mathcal{C}_T$ for each task T such that all $\boldsymbol{c} \in \mathcal{C}_T \subset \mathcal{C}$ describe T. Given a base language instruction $\boldsymbol{c}_{0,T} \in \mathcal{C}_T$, our goal is to generate a diverse set of N new instructions $\{\boldsymbol{c}_{1,T}, \ldots, \boldsymbol{c}_{N,T}\} \subset \mathcal{C}_T$. We evaluate the efficacy of the new instructions via several diversity metrics as well as the performance of the VLA after fine-tuning with these additional instructions.

TABLE V: Categories of failure in task instructions, partially obtained from [15]. For ease of reference, we index the attack styles using "z0" through "z7". Refer to Srikanth et al. [37] for details on how these failure modes were selected.

Step-by-step instructions: unnecessary breakdown of the task into multiple sequential steps **(z0)**	**Uncommon vocabulary:** use of rare, technical, or overly formal words instead of simple ones **(z1)**
Human-centric tone: instruction phrased as if addressing a human instead of a robot **(z2)**	**Use of adverbs:** adding vague modifiers like "carefully" or "gently" that don't add clarity **(z3)**
Unnecessary action specification: extra details about how to execute the task, beyond the core instruction **(z4)**	**Overly verbose reformulation:** inflated or wordy phrasing without changing the meaning **(z5)**
Colloquial/casual style: use of slang or conversational tone inappropriate for robotic tasks **(z6)**	**Mixed modality references:** adding sensory references (e.g., sight, sound) that the robot may not have **(z7)**



(a) BERT Diversity of prompt set.

(b) Failure Variance.

(c) Distance to original prompt.

Fig. 3: Diversity of our generated data compared to the Rephrase and ERT [15] baselines on OpenVLA-OFT. Each experiment was repeated 4 times, with error bars representing the standard error of the measurements. "Variance of Failure" is rescaled to be between 0 and 1. "Distance to Original" represents the average sentence embedding dissimilarity (1 - cosine similarity) for each domain. Q-DIG obtains the highest diversity metric in all cases.