Humanoid Everyday: A Comprehensive Robotic Dataset for Open-World Humanoid Manipulation

Zhenyu Zhao[1*] Hongyi Jing[1*] Xiawei Liu[1] Jiageng Mao[1†]
Abha Jha[1] Hanwen Yang[1] Rong Xue[1] Sergey Zakharov[2] Vitor Guizilini[2] Yue Wang[1†]
[1]University of Southern California [2]Toyota Research Institute

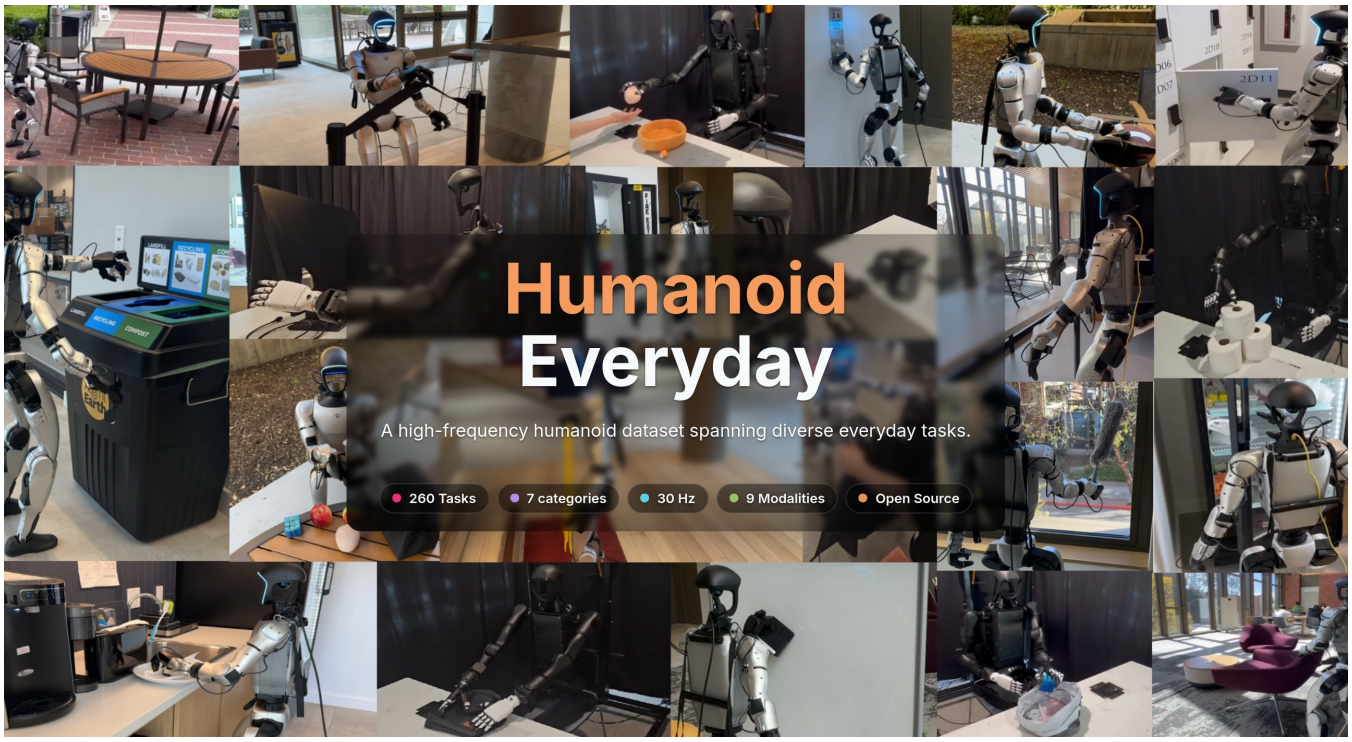


Fig. 1: **Humanoid Everyday**, a large-scale and diverse humanoid manipulation dataset characterized by extensive task varieties across a wide range of environments.

***Abstract*— From loco-motion to dextrous manipulation, humanoid robots have made remarkable strides in demonstrating complex full-body capabilities. However, the majority of current robot learning datasets and benchmarks mainly focus on stationary robot arms, and the few existing humanoid datasets are either confined to fixed environments or limited in task diversity, often lacking human-humanoid interaction and lower-body locomotion. Moreover, there are a few standardized evaluation platforms for benchmarking learning-based policies on humanoid data. In this work, we present Humanoid Everyday, a large-scale and diverse humanoid manipulation dataset characterized by extensive task variety involving dextrous object manipulation, human-humanoid interaction, locomotion-integrated actions, and more. Leveraging a highly efficient human-supervised teleoperation pipeline, Humanoid Everyday aggregates high-quality multimodal sensory data—including RGB, depth, LiDAR, and tactile inputs—together with natural language annotations, comprising 10.3k trajectories and over 3 million frames of data across 260 tasks across 7 broad categories. In addition, we conduct an analysis of representative policy learning methods on our dataset, providing insights into their strengths and limitations across different task categories. For standardized evaluation, we introduce a cloud-based evaluation platform that allows researchers to seamlessly deploy their policies in our controlled setting and receive performance feedback. By releasing Humanoid Everyday along with our policy learning analysis and a standardized cloud-based evaluation platform, we intend to advance research in general-purpose humanoid manipulation and lay the groundwork for more capable and embodied robotic agents in real-world scenarios. Our dataset, data collection code, and cloud evaluation website are made publicly available on our project website: `https://humanoideveryday.github.io`**

* Equal contribution. † Equal advising.

I. Introduction

Recent progress in humanoid robotics has significantly reduced the embodiment gap, enabling robots to perform dynamic activities such as running, dancing, and complex full-body movements [1]–[4]. However, collecting humanoid

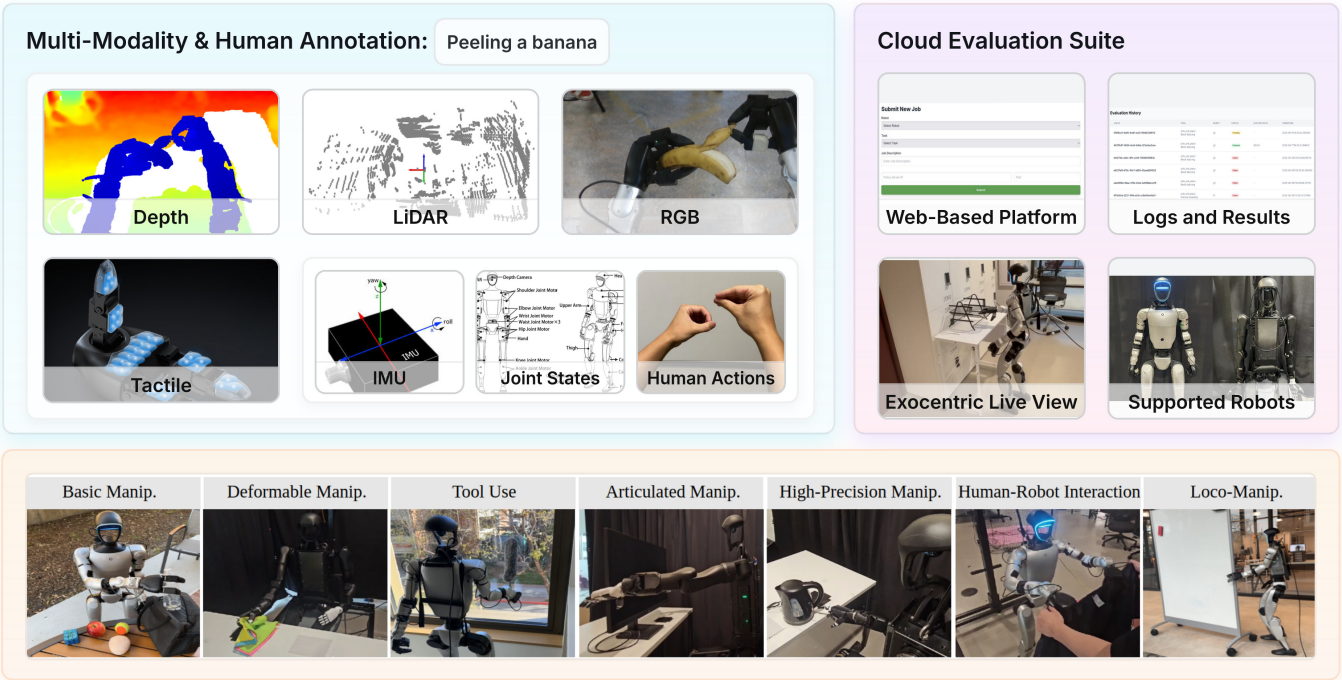


Fig. 2: **Humanoid Everyday Overview.** Humanoid Everyday covers 7 distinct categories of humanoid manipulation tasks with rich multimodal information, and provides a cloud-based evaluation platform for standardized policy deployment.

manipulation datasets remains challenging. It requires operating in both indoor and outdoor environments, executing a wide range of tasks, and leveraging the humanoid form through bimanual coordination, full-body motion, and human-centric interactions. Existing datasets mainly target stationary arms or mobile platforms with simple grippers and wheeled bases [5]–[9] (see Table I); even egocentric efforts like Humanoid Policy [10] emphasize repetitive tasks with limited locomotion. These limitations underscore the need for a diverse, interactive humanoid manipulation dataset that captures the full spectrum of human-like capabilities across varied environments and task complexities.

Beyond the scarcity of diverse and interactive datasets, there is also a notable lack of standardized evaluation practices for humanoid manipulation. Although recent deep learning methods have demonstrated improving performance in robotic manipulation [11]–[14], there remains a need for a unified evaluation framework for systematically comparing these approaches in the context of humanoid tasks. The absence of an evaluation standard makes it challenging to conduct fair and rigorous comparisons across policies, limiting our understanding of what truly drives effective humanoid manipulation in diverse scenarios. Together, these limitations lead to a ponderous question: what should an effective humanoid manipulation dataset look like to bridge the gaps in diversity, embodiment, and evaluation in the development of more intelligent robotic agents?

In this work, we introduce the Humanoid Everyday Dataset, a large-scale collection of 260 tasks across 7 categories (see Figure 2) that captures full-body locomotion, dexterous manipulation, and rich human-humanoid interactions in diverse environments. Unlike prior datasets limited in scope or simple settings [8]–[10], we accumulate more comprehensive and human-like tasks across various environments, retaining human supervision for data accuracy. Our additional technical re-engineering of the Unitree official teleoperation script enables sub-millisecond data synchronization across different modalities. At 30Hz, our pipeline captures high-resolution sensor and action data from every humanoid episode. We record each task with egocentric RGB video, depth information, LiDAR scans, tactile and Inertia Measurement Unit (IMU) information, joint poses, joint actions, and natural language annotations.

Beyond data collection, we analyzed representative policy learning methods for Humanoid Everyday, highlighting strengths and limitations across different task categories. These analyses provide initial insights into the challenges of embodied humanoid manipulation. To further support the community, we introduce a cloud-based evaluation platform that allows researchers to upload their policies, execute them within our standardized real-world environment, and receive detailed performance feedback. Unlike existing cloud evaluation systems that focus on robotic arms [15] or simulated agents [16], ours is the first platform designed for humanoid, which aims to lower the barrier to fair comparison and fosters collaborative progress in humanoid robotics. Together, Humanoid Everyday and our evaluation system offer a valuable foundation for advancing research on general-purpose humanoid manipulation for developing more robust, capable, and embodied agents in real-world scenarios.

Our contributions are threefold: (1) a large-scale multi-modal humanoid manipulation dataset collected in diverse

real-world scenarios with an optimized teleoperation pipeline; (2) an analysis of representative policy learning methods on Humanoid Everyday, highlighting their strengths and limitations across task categories; and (3) a cloud-based evaluation platform that enables standardized, reproducible, and collaborative research in humanoid manipulation.

Dataset	# Trajectories	# Skills	Lidar Sensing	Humanoid Robot	Dexterous Hand	Bipedal Loco-Manipulation	Human-Robot Interaction	Cloud Evaluation
MIME [17]	8.3k	12	✗	✗	✗	✗	✗	✗
RoboTurk [18]	2.1k	2	✗	✗	✗	✗	✗	✗
RoboNet [19]	162k	N/A	✗	✗	✗	✗	✗	✗
BridgeData [20]	7.2k	4	✗	✗	✗	✗	✗	✗
RH20T [6]	13k	33	✗	✗	✗	✗	✗	✗
DROID [21]	76k	86	✗	✗	✗	✗	✗	✗
Open X-Embodiment [5]	1.4M	217	✗	✗	✗	✗	✗	✗
Fourier ActionNet [8]	13k	16	✗	✓	✓	✗	✗	✗
Agibot World [9]	1M	87	✗	✗*	✗†	✗	✓	✗
Humanoid Everyday	10.3k	221	✓	✓	✓	✓	✓	✓

* Wheeled base instead of bipedal legs, not considered a humanoid robot here. † Mostly grippers and only a few dexterous hands.

TABLE I: Comparison of Robotic Manipulation Datasets

II. Related Work

A. Robotic Manipulation Datasets

Robot manipulation datasets have played an important role in enabling data-driven robot policy learning in recent years. Some datasets focus on a single pattern of behavior, such as pushing [19], [22], grasping [23], [24], and pouring [25], [26]. Some other large-scale multi-task datasets have also been proposed to improve generalization across diverse manipulation scenarios and environmental conditions, but still focus on robotic arm platforms [5], [7], [27]. More recently, the rise of humanoid robots has led to the creation of humanoid-specific datasets. [8], [10], [28] leverage motion capture data or human teleoperation to train humanoid agents in either simulated or constrained tabletop settings, focusing on simplified manipulation tasks. [9], [29] collect real-world humanoid robotic data, demonstrating diverse manipulation skills using upper-body teleoperation. Despite these advances, most existing datasets focus on robotic-arm manipulation or humanoids with limited capabilities, often lacking diverse tasks, complex environments, and full-body functions. In contrast, Humanoid Everyday captures a broad spectrum of full-body humanoid activities across various indoor and outdoor settings, offering a comprehensive resource for general-purpose humanoid policy development.

B. Humanoid Robot Learning

Recent works have explored diverse learning-based policies on humanoid robots. Many of them leverage reinforcement learning and sim-to-real transfer to achieve whole-body coordination for balance and locomotion [1], [30]–[34], some others also incorporate human motion datasets in the RL training pipeline to ease sim-to-real transfer [2], [35]. Beyond whole-body control, learning humanoid manipulation has also become an active research direction. Many approaches begin by collecting demonstration data through teleoperation using VR applications [36]–[39]. Imitation learning and Vision-Language-Action (VLA) models are then utilized to train corresponding policies, allowing robots to replicate demonstrated behaviors with high fidelity [2], [37], [38], [40]–[44]. However, current approaches are inherently limited by the diversity of collected demonstrations, which mostly concentrate on a narrow set of tasks or environments. Consequently, few works have systematically evaluated humanoid policies across diverse task categories. Humanoid Everyday offers the diversity and scale needed to train more robust and generalizable manipulation policies, and our analysis of representative imitation learning approaches provides insights into their strengths and limitations, thereby facilitating more efficient learning and execution of complex behaviors.

C. Robot Policy Evaluation

Beyond the availability of humanoid datasets and training methods, fair evaluation of robotic policies is critical for measuring progress and ensuring reproducibility across diverse systems and environments. [40], [45]–[47] develop benchmarking suites and evaluation protocols that assess robotic performance in simulation systems under controlled conditions. In addition to simulation-based evaluation, directly assessing policies in real-world settings is also a common practice. [48]–[50] have proposed simple and reproducible real-world robotic setups to enable consistent policy inference and evaluation under similar environments. AutoEval [15] allows users to upload and deploy different learned policies under the same standardized setting for autonomous evaluation. However, these evaluation frameworks are largely confined to robotic arm platforms and do not extend to humanoid robots with their unique challenges. To fill this gap, Humanoid Everyday introduces a cloud-based evaluation platform tailored for humanoid robots. Our platform enables deployment on a standardized real-world humanoid system, providing consistent inference and evaluation for benchmarking manipulation policies across diverse tasks and

users.

III. Humanoid Everyday Dataset

To support learning and evaluation of humanoid manipulation across various real-world tasks, we introduce the Humanoid Everyday Dataset, a large-scale, high-quality dataset collected with full-body humanoid robots, spanning diverse tasks including human-humanoid interaction and loco-manipulation. In Section III-A, we describe the hardware setup, including the humanoid platforms and the operation interface used for data collection. Section III-B outlines our data collection pipeline and describes how our method improves the overall efficiency. Finally, we present the composition and structure of the dataset in Section III-C.

A. Environment setup

a) Hardware: We gather data with two Unitree humanoid robots: the 29-degrees of freedom (DoF) **G1** with 7-DoF three-fingered dexterous hands (Dex3-1) and the 27-DoF **H1** with 6-DoF INSPIRE hands. Both the H1 and G1 humanoid robots are equipped with Intel RealSense RGB-D cameras and a Livox LiDAR system. In addition, the G1's Dex3-1 hands possess tactile sensors, which further enhances the multimodality of our dataset.

b) Teleoperation interface: The operator wears an Apple Vision Pro to capture wrist and finger keypoints using its cameras on the bottom. The finger motions are mapped onto the robot's dexterous hands via the dex-retargeting system [51] for the robot hands to complete basic manipulations. The wrist poses are converted into arm joint commands through a Pinocchio-based inverse kinematics algorithm [52], enabling full upper-body teleoperation.

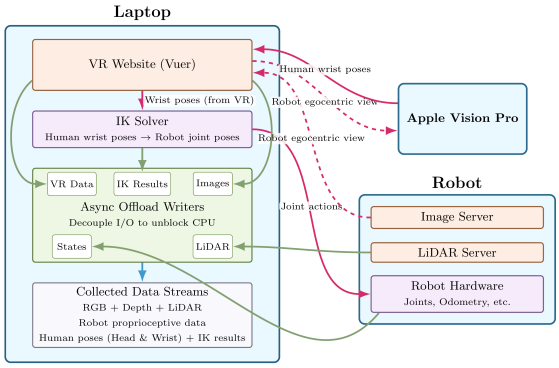
B. Efficient and Scalable Data Collection

We propose a multi-processing teleoperation pipeline, re-engineered upon the official Unitree teleoperation library and significantly improved for large-scale, high-quality humanoid data collection. Our intuitive and robust design enables low-latency teleoperation with high-frequency control while ensuring synchronized, high-quality data streams.

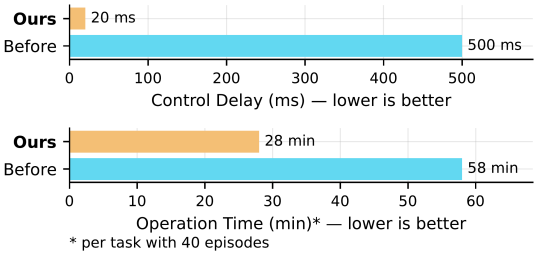
As opposed to the blocking and synchronous design in the Unitree official teleoperation script, our data collection pipeline uses multi-processing and asynchronous IO reading and writing to ensure high-frequency teleoperation and high-quality data collection. Specifically, as shown in Figure 3a, we decouple IO data from inverse kinematics (IK) computation and robot joint control in separate processes, with shared memory buffers facilitating fast and low-latency inter-process communication. This design allocates more computational resources to the IK solver, enabling smoother and higher-frequency teleoperation control. In addition, data recording and sensor processing are handled asynchronously in parallel threads within main pipeline process, ensuring nonblocking and temporally aligned data collection.

Additionally, we provide a simple data-collection interface that hides system complexity, stream the robot's binocular IR feeds to the VR headset for better situational awareness, and support multiple recording sessions per run without restarting the entire program.

All data collection was performed on a laptop equipped with an 11th Gen Intel i7 CPU. As shown in Figure 3b, our pipeline halves data collection time compared to the official Unitree teleoperation system, while the control delay decreases from 500 ms to 2 ms. These improvements highlight the efficiency and scalability of our pipeline, enabling rapid and high-quality data acquisition for complex humanoid tasks.



(a)



(b)

Fig. 3: **Data Collection Pipeline.** (a) We separate data streaming, data writing, robot control, and IK computation into distinct processes and threads to ensure reliable and efficient data collection. (b) Our pipeline substantially reduces control delay and enhances data collection efficiency.

C. A Diverse Collection of Everyday Humanoid Tasks

Humanoid Everyday is a large-scale, diverse collection of humanoid manipulation tasks, comprising seven main distinct categories including Basic Manipulation (basic object pick-and-place manipulation), Deformable Manipulation (interacting with cloths or other deformable objects), Articulated Manipulation (operating hinged or jointed structures), Tool Use (utilizing external objects to achieve goals), High-Precision Manipulation (performing difficult tasks requiring high accuracy), Human-Robot Interaction (engaging in cooperative actions with humans), and Loco-Manipulation (combining locomotion and manipulation) (see Figure 4. Our tasks are performed in indoor and outdoor environments, involving complex interactions with surrounding objects and dynamic settings. Additionally, a portion of these tasks require lower-body locomotion, adding further diversity to the dataset.

This variety in environmental contexts enriches the collected data and supports the development of generalizable policies capable of adapting to diverse real-world scenarios.

Humanoid Everyday comprises 260 unique tasks, each with 40 episodes to provide sufficient training data. We record each task with a rich set of sensory modalities, offering a comprehensive view of humanoid interactions. Each episode includes RGB video, depth maps, LiDAR, tactile feedback, and natural language task descriptions. This multimodal design enables richer training trajectories, supporting the development of more adaptable and context-aware humanoid policies.

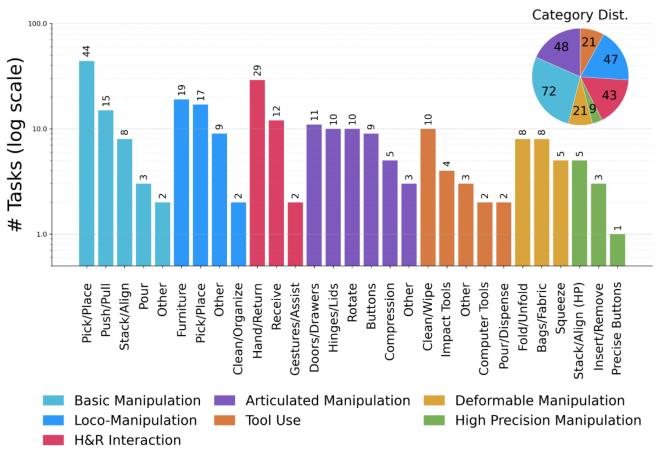


Fig. 4: **Data Distribution.** Distribution of tasks and skill categories in the Humanoid Everyday Dataset.

IV. Evaluating Humanoid Manipulation Policies

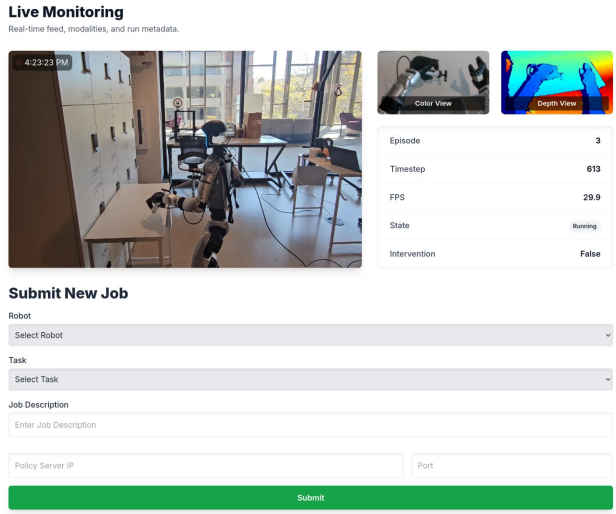


Fig. 5: **Evaluation platform.** Humanoid Everyday introduces a cloud-based evaluation platform on real-world humanoid setups

To enable systematic and reproducible evaluation of humanoid manipulation policies, we introduce a cloud-based evaluation platform tailored for humanoid robots, as shown in Figure 5.

Our cloud-based evaluation platform enables remote policy deployment on locally hosted humanoid robots, addressing the hardware access bottleneck in real-world policy inference. By allowing researchers to test their policies on actual humanoid robots without owning them, our platform significantly lowers the barrier to humanoid manipulation learning research and standardizes the evaluation process across different methods and users.

We reconstruct a subset of task environments from the Humanoid Everyday Dataset to support realistic and standardized evaluations. Researchers wishing to evaluate their trained policies over the Humanoid Everyday Dataset can connect to the platform by simply specifying their policy server IP and port, after which our system streams real-time visual inputs (RGB images and Depth information) and robot state information from the G1 or H1 robot to the client. Users can then run inference using their own policies locally and send the resulting action commands back to our server. These commands are executed in real time on a physical humanoid robot. In addition, we stream both egocentric RGB and depth images from the robot and third-person camera views to a web interface, allowing users to monitor the task execution remotely.

Our evaluation platform framework enables highly efficient deployment of humanoid policies. As shown in Figure 6, our online evaluation system runs continuously for over 100 minutes before battery depletion. Only three human interventions were required due to motor overheating, and the system otherwise maintained high evaluation efficiency throughout the rest of the process. This cloud-based system enables reproducible and hardware-agnostic evaluation of humanoid manipulation policies on our Humanoid Everyday dataset. We believe it could provide a practical solution for researchers without direct access to humanoid robots, and could establish a common testbed for fair benchmarking and collaborative development.

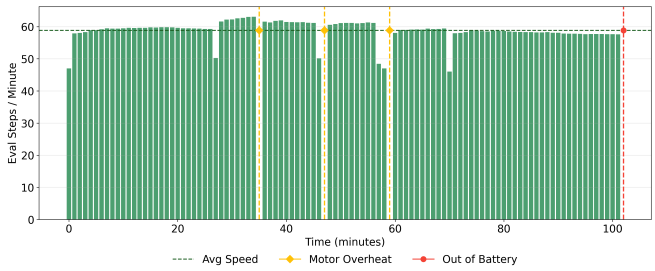


Fig. 6: Evaluation Steps per Minute with Human Interventions. Our evaluation runs continuously for over 100 minutes before running out of battery, while only three human interventions were required due to motor overheating. The system maintained consistently high evaluation efficiency throughout the rest of the process.

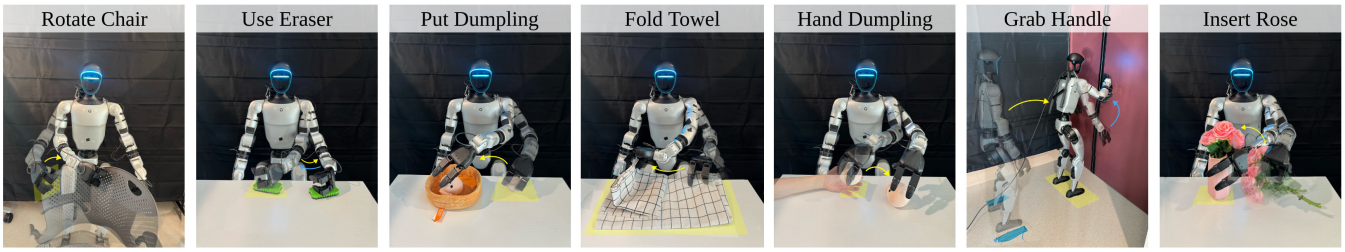


Fig. 7: **Experiment Setup**. Representative inference tasks from the seven categories. The yellow region denotes the task execution area with slight variations, and arrows show the robot's arm trajectories. For each task, we conduct 10 trials under seven different policies.

V. Experiments

A. Performance of Imitation Learning Policies

To better understand how existing imitation learning approaches perform in the context of humanoid manipulation, we evaluate a variety of policies on our Humanoid Everyday dataset, including Diffusion Policy (DP) [11], 3D Diffusion Policy (DP3) [12], Action Chunking with Transformers (ACT) [13], OpenVLA [14], π_0-FAST [53], $\pi_{0.5}$ [54], and GR00T N1.5 [55]. We train these policies on our 30Hz humanoid data until convergence. For VLA-based policies, we adopt a two-stage finetuning strategy: we first fine-tune on the full Humanoid Everyday dataset and then further adapt each model to task-specific data from individual categories.

We conduct experiments across all seven task categories in Humanoid Everyday. For detailed per task experiment setup, see Figure 7. We present the results of these policies in Table II. where each policy is evaluated over 10 trials on seven tasks.

Overall, all the end-to-end imitation policies struggle in humanoid manipulation tasks due to the high-dimensional action space in our dataset (which is 28 DoFs in total). Among them, DP3 outperforms DP in most cases, and we hypothesize that the use of 3D point cloud observations provides stronger robustness to environmental variations. However, in Loco-Manipulation tasks where the robot must move and the point cloud undergoes large frame-to-frame changes, 3D-based inputs become less reliable than RGB images, leading to failure cases. ACT, on the other hand, performs poorly overall, as it does not effectively incorporate visual feedback and tends to overfit to demonstrated trajectories, mechanically replaying them without adapting to scene variations.

In contrast, large VLA models demonstrate more consistent and stable performance on humanoid manipulation, benefiting from pretraining priors that improve generalization, particularly on Deformable Manipulation and Loco-Manipulation tasks that require both precision and robustness. OpenVLA does not compress the action space and thus when trained on high-frequency 30 Hz data, it often fails to produce meaningful motions. Downsampling to 2 Hz alleviates this issue, but the resulting behaviors appear less smooth. π_0-FAST employs a DCT-based tokenizer for action compression, but the high-dimensional humanoid actions are not well represented, leading to excessive token numbers, incorrect output tokens, and decoding errors that sometimes cause the robot to halt during inference. On the other hand, although $\pi_{0.5}$ produces smoother humanoid actions compared to the previous two models, it still tends to overfit to trajectories, ignoring visual feedback. In particular, GR00T N1.5 achieves the strongest performance overall, largely due to its extensive pretraining across multiple large-scale humanoid datasets, which provides a powerful prior well-suited for our diverse manipulation tasks.

Despite these differences, all imitation learning policies struggle significantly on the most challenging categories, such as Loco-Manipulation and High-Precision Manipulation tasks. In the "Insert rose into vase" task, nearly all policies achieve a 0% success rate: while many can lift the rose, they consistently fail to insert its thin stem into the vase, suggesting that current models lack fine-grained visuospatial perception. This indicates that imitation learning on humanoid platforms still faces significant challenges and opportunities for progress.

Task Category	**Task**	**DP**	**DP3**	**ACT**	**OpenVLA**	**π_0-FAST**	**$\pi_{0.5}$**	**GR00T N1.5**
Articulate	Rotate chair	100%	90%	100%	70%	100%	100%	100%
Tool Use	Use eraser to wipe the desk	0%	70%	0%	30%	40%	40%	0%
Basic	Put dumpling toy into plate	30%	20%	70%	30%	60%	30%	80%
Deformable	Fold towel on the desk	0%	20%	0%	40%	20%	40%	50%
HRI	Hand over dumpling toy	40%	40%	70%	60%	30%	40%	100%
Loco-Manip.	Walk to grab door handle	30%	0%	0%	30%	10%	0%	30%
High Precision	Insert rose into vase	0%	0%	0%	10%	0%	0%	0%
Average		29%	34%	34%	39%	37%	36%	51%

TABLE II: Success rates of imitation learning methods on the Humanoid Everyday Dataset.

B. Pretraining on Humanoid Everyday as a Prior

To validate whether Humanoid Everyday can serve as an effective pretraining prior for large Vision-Language-Action (VLA) models, we design an ablation study comparing two training strategies: (i) the two-stage finetuning pipeline introduced in the previous section, and (ii) direct task-specific finetuning without pretraining on Humanoid Everyday. We select the Human–Robot Interaction task of handing over a dumpling toy, which represents a moderately challenging manipulation scenario. This task requires both coordination and robustness to interaction dynamics, yet is tractable enough for all VLA models to achieve a non-trivial level of success.

As shown in Figure 8, finetuning on Humanoid Everyday prior to task-specific adaptation consistently improves the performance of VLA models. This suggests that exposure to diverse humanoid behaviors provides a useful prior that facilitates learning downstream manipulation tasks, indicating that large-scale and diverse humanoid data can improve robustness and stability in humanoid manipulation.

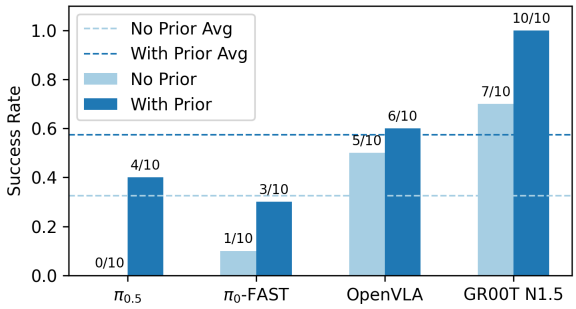


Fig. 8: **Ablation on Humanoid Everyday Pretraining.** Performance comparison between direct task-specific finetuning and two-stage finetuning with Humanoid Everyday.

VI. Discussion and Conclusion

We introduce Humanoid Everyday, a large-scale, diverse humanoid manipulation dataset encompassing full-body locomotion, dexterous manipulation, and rich human-humanoid interactions in various everyday settings. Leveraging this dataset, we conduct policy learning analyses that highlight both the strengths and shortcomings of current methods in humanoid manipulation. Additionally, we provide a cloud-based evaluation platform that enables policies to be deployed directly on our humanoid robot settings, promoting reproducibility and collaborative progress within the global humanoid robotics community. We believe Humanoid Everyday will serve as a valuable resource for advancing versatile and intelligent humanoid robotic agents.

While Humanoid Everyday provides a comprehensive dataset for humanoid manipulation, we only evaluate existing imitation learning policy architectures. Although these baselines provide useful insights, their performance degrades on more challenging tasks due to the high dimensionality of humanoid action spaces, indicating the need for more specialized model designs. In addition, our cloud-based evaluation system does not yet support automatic scene resetting, as current imitation learning policies are not sufficiently robust for humanoids to recover the environments without human assistance. In future work, we plan to develop more robust humanoid policies and extend our evaluation system to support autonomous scene recovery.

VII. Acknowledgements

The USC Physical Super Intelligence Lab acknowledges generous supports from Toyota Research Institute, Dolby, Google DeepMind, Capital One, Nvidia, and Qualcomm. Yue Wang is also supported by a Powell Research Award.

References

[1] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[2] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[3] Y. Xue, W. Dong, M. Liu, W. Zhang, and J. Pang, "A unified and general humanoid whole-body controller for fine-grained locomotion," *arXiv preprint arXiv:2502.03206*, 2025.

[4] J. Mao, S. Zhao, S. Song, T. Shi, J. Ye, M. Zhang, H. Geng, J. Malik, V. Guizilini, and Y. Wang, "Learning from massive human videos for universal humanoid pose control," *arXiv preprint arXiv:2412.14172*, 2024.

[5] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[6] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," *arXiv preprint arXiv:2307.00595*, 2023.

[7] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[8] Y. M. Fourier ActionNet Team, "Actionnet: A dataset for dexterous bimanual manipulation," 2025.

[9] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Huang, S. Jiang *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[10] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, G. Yang, J. Zhang, S. Yi, G. Shi, and X. Wang, "Humanoid policy ˜ human policy," *arXiv preprint arXiv:2503.13441*, 2025.

[11] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[12] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[13] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[14] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[15] Z. Zhou, P. Atreya, Y. L. Tan, K. Pertsch, and S. Levine, "Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world," *arXiv preprint arXiv:2503.24278*, 2025.

[16] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani *et al.*, "Evaluating real-world robot manipulation policies in simulation," *arXiv preprint arXiv:2405.05941*, 2024.

[17] P. Sharma, L. Mohan, L. Pinto, and A. Gupta, "Multiple interactions made easy (mime): Large scale demonstrations data for imitation," in *Conference on robot learning*. PMLR, 2018, pp. 906–915.

[18] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*. PMLR, 2018, pp. 879–893.

[19] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[20] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," *arXiv preprint arXiv:2109.13396*, 2021.

[21] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[22] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2786–2793.

[23] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3511–3516.

[24] S. Zhao, J. Mao, W. Chow, Z. Shangguan, T. Shi, R. Xue, Y. Zheng, Y. Weng, Y. You, D. Seita *et al.*, "Robot learning from any images," *arXiv preprint arXiv:2509.22970*, 2025.

[25] H. Lin, Y. Fu, and X. Xue, "Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 241–251.

[26] A. Burns, S. Xiang, D. Lee, L. Jackel, S. Song, and V. Isler, "Look and listen: A multi-sensory pouring network and dataset for granular media from human demonstrations," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2519–2524.

[27] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju *et al.*, "Robocat: A self-improving foundation agent for robotic manipulation," *arXiv preprint arXiv:2306.11706*, vol. 1, no. 8, 2023.

[28] N. Wagener, A. Kolobov, F. Vieira Frujeri, R. Loynd, C.-A. Cheng, and M. Hausknecht, "Mocapact: A multi-task dataset for simulated humanoid control," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 418–35 431, 2022.

[29] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with improved 3d diffusion policies," *arXiv preprint arXiv:2410.10803*, 2024.

[30] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through multi-task reinforcement learning," *arXiv preprint arXiv:2302.09450*, vol. 1, 2023.

[31] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik, "Humanoid locomotion as next token prediction," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[32] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu, "Deep imitation learning for humanoid loco-manipulation through human teleoperation," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2023, pp. 1–8.

[33] A. Tang, T. Hiraoka, N. Hiraoka, F. Shi, K. Kawaharazuka, K. Kojima, K. Okada, and M. Inaba, "Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 107–13 114.

[34] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.

[35] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[36] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[37] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[38] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8944–8951.

[39] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang, "Mobile-television: Predictive motion priors for humanoid whole-body control," *arXiv preprint arXiv:2412.07773*, 2024.

[40] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.

[41] O. Mees, L. Hermann, and W. Burgard, "What matters in language conditioned robotic imitation learning over unstructured data," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 205–11 212, 2022.

[42] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–20, 2021.

[43] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[44] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, "Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control," *arXiv preprint arXiv:2505.03738*, 2025.

[45] Y. Lee, E. S. Hu, and J. J. Lim, "Ikea furniture assembly environment for long-horizon complex manipulation tasks," in *2021 ieee international conference on robotics and automation (icra)*. IEEE, 2021, pp. 6343–6349.

[46] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[47] Y. R. Wang, C. Ung, G. Tannert, J. Duan, J. Li, A. Le, R. Oswal, M. Grotz, W. Pumacay, Y. Deng *et al.*, "Roboeval: Where robotic manipulation meets structured and scalable evaluation," *arXiv preprint arXiv:2507.00435*, 2025.

[48] B. Yang, J. Zhang, V. Pong, S. Levine, and D. Jayaraman, "Replab: A reproducible low-cost arm benchmark platform for robotic learning," *arXiv preprint arXiv:1905.07447*, 2019.

[49] M. Heo, Y. Lee, D. Lee, and J. J. Lim, "Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation," *The International Journal of Robotics Research*, p. 02783649241304789, 2023.

[50] J. Luo, C. Xu, F. Liu, L. Tan, Z. Lin, J. Wu, P. Abbeel, and S. Levine, "Fmb: a functional manipulation benchmark for generalizable robotic learning," *The International Journal of Robotics Research*, vol. 44, no. 4, pp. 592–606, 2025.

[51] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *arXiv preprint arXiv:2307.04577*, 2023.

[52] J. Carpentier, F. Valenza, N. Mansard *et al.*, "Pinocchio: fast forward and inverse dynamics for poly-articulated systems, 2015–2018," *URL https://stack-of-tasks. github. io/pinocchio*.

[53] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

[54] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[55] J. Bjorck, V. Blukis, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, X. Jiang, K. Kundalia, J. Kautz, Z. Li, Z. Lin, K. Lin, L. Magne, Y. Man, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, J. Wang, Q. Wang, S. Wang, J. Xiang, Y. Xie, Y. Xu, S. Ye, Z. Yu, Y. Zhao, Z. Zhang, R. Zheng, and Y. Zhu, "Gr00t n1.5: An improved open foundation model for generalist humanoid robots," Model Card / Technical Report, NVIDIA Research, June 2025, available: https://research.nvidia.com/labs/gear/gr00t-n1_5/.