Generalizable Humanoid Manipulation with Improved 3D Diffusion Policies

Yanjie Ze[1] Zixuan Chen[2] Wenhao Wang[3] Tianyi Chen[3]
Xialin He[4] Ying Yuan[5] Xue Bin Peng[2] Jiajun Wu[1]
[1]Stanford University [2]Simon Fraser University [3]UPenn [4]UIUC [5]CMU

HUMANOID-MANIPULATION.GITHUB.IO

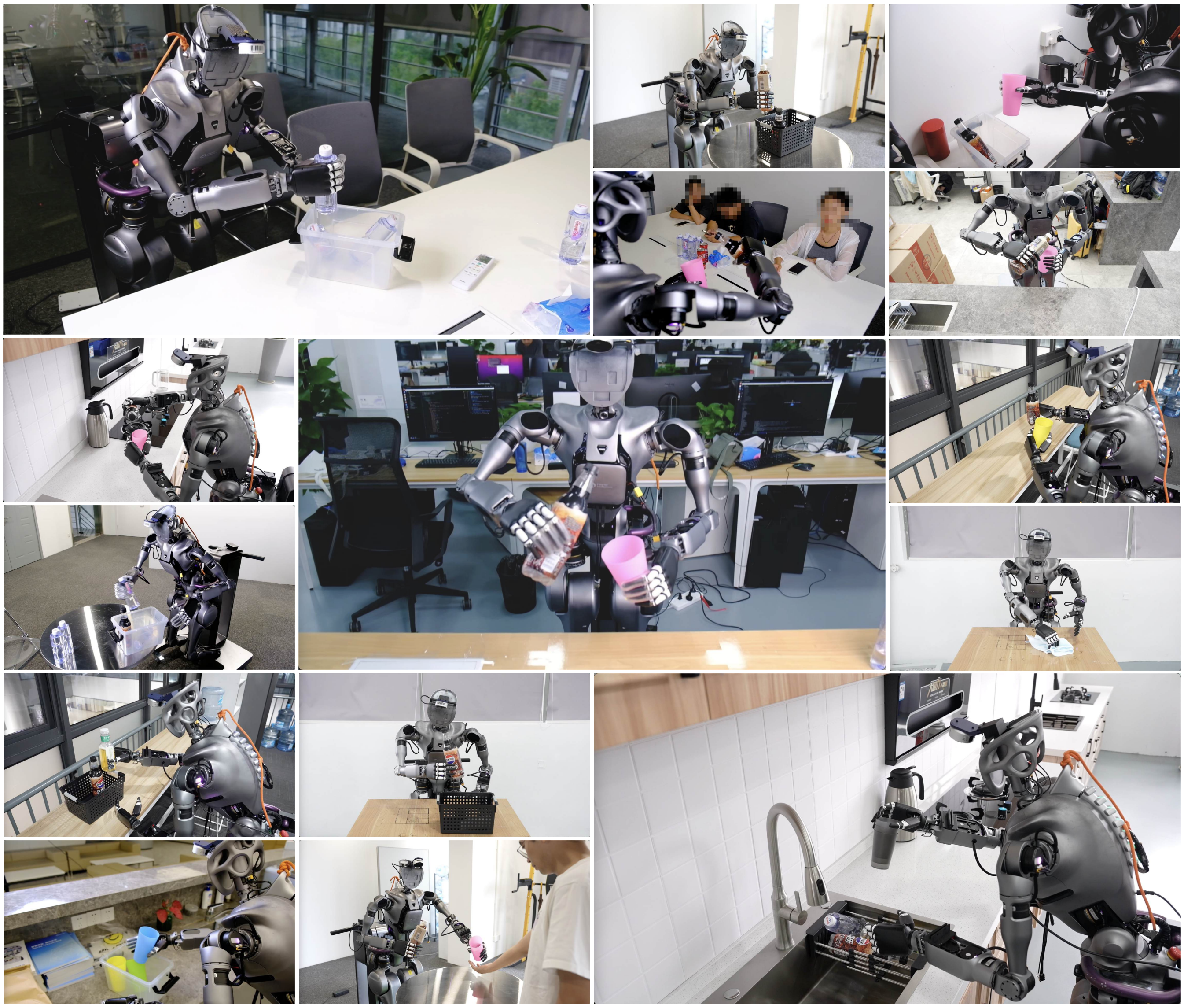

Fig. 1: **Humanoid manipulation in diverse unseen scenarios.** With only data collected from a single scene, our Improved 3D Diffusion Policy (iDP3) enables a full-sized humanoid robot to perform practical skills in diverse real-world environments. **The scenes are not cherry-picked.** Videos are available on our website.

Abstract— **Humanoid robots capable of autonomous operation in diverse environments have long been a goal for roboticists. However, autonomous manipulation by humanoid robots has largely been restricted to one specific scene, primarily due to the difficulty of acquiring generalizable skills. Recent advances in 3D visuomotor policies, such as the 3D Diffusion Policy (DP3), have shown promise in extending these capabilities to wilder environments. However, 3D visuomotor policies often rely on camera calibration and point-cloud segmentation, which present challenges for deployment on mobile robots like humanoids. In this work, we introduce the Improved 3D Diffusion Policy (iDP3), a novel 3D visuomotor policy that eliminates these**

constraints by leveraging egocentric 3D visual representations. We demonstrate that iDP3 enables a full-sized humanoid robot to autonomously perform skills in diverse real-world scenarios, using only data collected in the lab. Videos are available at humanoid-manipulation.github.io.

I. Introduction

Humanoid robots capable of performing diverse tasks in unstructured environments have long been a significant goal in the robotics community. Recently, there has been substantial progress in the development of humanoid robot hardware [1]–[5]. Simultaneously, visual imitation learning methods for controlling these robots have gained popularity, enabling them to autonomously execute complex skills [6]–[11]. However, most of these autonomous manipulation skills are still largely confined to a specific scenario [6]–[11], mainly due to the restricted generalization capabilities of visual imitation learning approaches [12]–[16].

Recent advances in 3D visuomotor policies have shown great potential to generalize the learned skills to more complex and diverse scenarios [17]–[21]. Among these, the 3D Diffusion Policy (DP3, [17]) is effective in a variety of simulated and real-world tasks across different embodiments. These include deformable object manipulation with a dexterous hand [17] or a mobile arm [22], long-horizon bi-manual manipulation [10], and loco-manipulation with a quadrupedal robot [23]. Despite DP3's generalizability, its applications have been restricted to tasks performed using a third-person view with a calibrated fixed camera, largely due to the need for accurate camera calibration and point-cloud segmentation, both of which are inherent challenges in 3D visuomotor policies.

In this work, we aim to develop generalizable humanoid robotic manipulation skills using 3D visuomotor policies. To address the limitations of existing 3D visuomotor policies for humanoid robots, we propose the **Improved 3D Diffusion Policy (iDP3)**, a novel 3D imitation learning method that leverages egocentric 3D representations in the camera frame, eliminating the need for camera calibration and point cloud segmentation. Additionally, we introduce several modifications to improve the effectiveness of iDP3 significantly.

For data collection, we design a whole-upper-body teleoperation system that maps human joints to a full-sized humanoid robot. Unlike the common bi-manual manipulation system, our teleoperation incorporates waist degrees of freedom and active vision, greatly expanding the robot's operational workspace, particularly when handling tasks at varying heights.

Through extensive real-world experiments and ablation studies, we demonstrate that iDP3 exhibits remarkable generalization across diverse scenes and shows strong view invariance, along with high effectiveness.

Our core contributions are summarized as follows:

- We introduce the Improved 3D Diffusion Policy (iDP3), a 3D visuomotor policy that can be applied to any robot, supporting both egocentric and third-person views, while achieving high efficiency and strong generalization abilities.
- We develop a whole-upper-body teleoperation system for a humanoid robot, enabling efficient data collection from humans.
- We demonstrate that our policy deployed on a humanoid robot can successfully generalize contact-rich manipulation skills to a wide range of real-world scenarios, with data collected in a single scene.

II. Related Work

A. Visuomotor Policy Learning

Classical approaches depend on state estimation to address robotic manipulation tasks [24]. Recently, there has been a growing trend in learning a visuomotor policy in an end-to-end manner to solve robotics problem [12], [17], [25]–[28]. There are two primary pathways: *imitation learning* [12], [15]–[21], [29]–[34] and *sim-to-real reinforcement learning* [35]–[44]. This work focuses on visual imitation learning, due to its strength in completing complex, diverse, and long-horizon tasks.

Image-based imitation learning methods, such as Diffusion Policy [12], have achieved significant success [10], [17], [22], [30], [45], while their limited generalization abilities restrict their application in complex real-world environments. Several recent works aim to address these limitations [17], [22], [45]–[47]. Among these, the 3D Diffusion Policy (DP3, [17]) has demonstrated notable generalization abilities and broad applicability to diverse robotic tasks [10], [11], [22], [23]. Nonetheless, 3D visuomotor policies are inherently dependent on precise camera calibration and fine-grained point cloud segmentation [17], [18], [21], [39], [47], which limits their deployment on mobile platforms such as humanoid robots. This work tackles this important problem and extends the application of 3D visuomotor policies into a more general setting.

Additionally, several recent works have demonstrated capabilities similar to ours. Maniwhere [37] achieves real-world scene generalization via large-scale simulation data. However, due to the significant sim-to-real gap, they only show tasks like pushing in unseen scenarios, rather than contact-rich tasks like pick and place. The Robot Utility Model [48] also generalizes skills to the new environment with imitation learning, while they have to use data collected from 20 scenes for scene generalization, compared to only 1 scene we use. VISTA [47] demonstrates impressive view generalization using view synthesis models. In contrast to their complex pipeline, we find that our egocentric 3D representations naturally enable robust view invariance.

B. Humanoid Robot Learning

The autonomous execution of diverse skills by humanoid robots in complex, real-world environments has long been a central goal in robotics. Recently, learning-based methods have shown promising progress toward this objective, particularly in the areas of locomotion [36], [49]–[52], manipulation [9], [11], [53], and loco-manipulation [6]–[8], [54]. While several works have successfully demonstrated

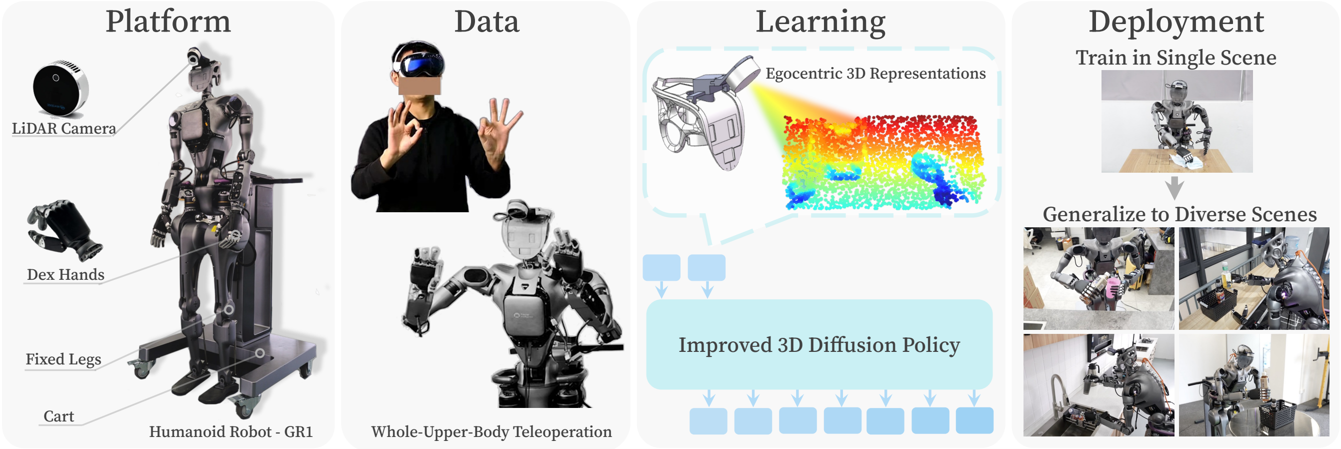
Platform
LiDAR Camera
Dex Hands
Fixed Legs
Cart
Humanoid Robot - GR1
Data
Whole-Upper-Body Teleoperation
Learning
Egocentric 3D Representations
Improved 3D Diffusion Policy
Deployment
Train in Single Scene
Generalize to Diverse Scenes

Fig. 2: **Overview of our system.** Our system mainly consists of four parts: the humanoid robot platform, the data collection system, the visuomotor policy learning method, and the real-world deployment. For the learning part, we develop Improved 3D Diffusion Policy (iDP3) as a visuomotor policy for general-purpose robots.

humanoid locomotion in unstructured, real-world environments [36], [49], [50], manipulation skills in unseen environments remain largely unexplored [6], [8], [9]. In this paper, we take a significant step forward by showcasing how the repurposed 3D visuomotor policy framework enables humanoid robots to perform manipulation tasks in unseen real-world scenes.

III. Improved 3D Diffusion Policy

3D Diffusion Policy (DP3, [17]) is an effective 3D visuomotor policy that marries sparse point cloud representations with diffusion policies. Although DP3 has shown impressive results across a wide range of manipulation tasks, it is not directly deployable on general-purpose robots such as humanoid robots or mobile manipulators due to its inherent dependency on precise camera calibration and fine-grained point cloud segmentation. Furthermore, the accuracy of DP3 requires further improvements for effective performance in more complex tasks. In the following, we detail several modifications to achieve targeted improvements. The resulting improved algorithm is termed as the **Improved 3D Diffusion Policy (iDP3)**.

Egocentric 3D Visual Representations. DP3 leverages a 3D visual representation in the world frame, enabling easy segmentation of the target object [17], [53]. However, for general-purpose robots like humanoids, the camera mount is not fixed, making camera calibration and point cloud segmentation impractical. To tackle this problem, we propose directly using the 3D representation from the camera frame, as shown in Figure 3. We term this class of 3D representations as *egocentric 3D visual representations*.

Scaling Up Vision Input. Leveraging egocentric 3D visual representations presents challenges in eliminating extraneous point clouds, such as backgrounds or tabletops, especially without relying on foundation models. To mitigate this, we propose a straightforward but effective solution: scaling up the vision input. Instead of using standard sparse point sampling as in previous systems [17], [22], [53], we significantly increase the number of sample points to capture the entire scene. Despite its simplicity, this approach proves to be effective in our real-world experiments.

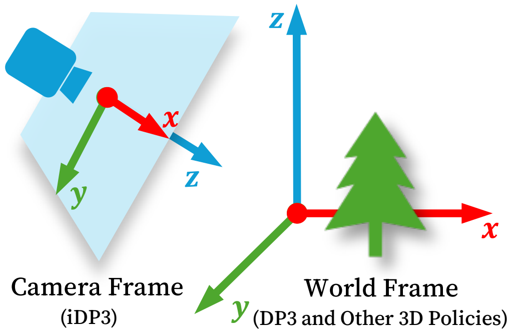

Fig. 3: **iDP3 utilizes 3D representations in the camera frame,** while the 3D representations of other recent 3D policies including DP3 [17] are in the world frame, which relies on accurate camera calibration and can not be extended to mobile robots.

Improved Visual Encoder. We replace the MLP visual encoder in DP3 with a pyramid convolutional encoder. We find that convolutional layers produce smoother behaviors than fully-connected layers when learning from human data, and incorporating pyramid features from different layers further enhances accuracy.

Longer Prediction Horizon. The jittering from human experts and the noisy sensors exhibit much difficulty in learning from human demonstrations, which causes DP3 to struggle with short-horizon predictions. By extending the prediction horizon, we effectively mitigate this issue.

Implementation Details. For the optimization, we train 300 epochs for iDP3 and all other methods with AdamW [55]. For the diffusion process, we use 50 training steps and 10 inference steps with DDIM [56]. For the point cloud sampling, we replace farthest point sampling (FPS) used in DP3 [17] with a cascade of voxel sampling and uniform sampling, which ensures the sampled points cover the 3D space with a faster inference speed.

IV. Humanoid Manipulation with Improved 3D Diffusion Policy

In this section, we present our real-world imitation learning system deployed on a full-sized humanoid robot. An overview of the system is provided in Figure 2.

A. Platform

Humanoid Robot. We use Fourier GR1 [5], a full-sized humanoid robot, equied with two Inspire Hands [57]. We enable the whole upper body {*head, waist, arms, hands*}, totaling 25 degrees-of-freedom (DoF). We disable the lower body for stability and use a cart for movement.

LiDAR Camera. To capture high-quality 3D point clouds, we utilize the RealSense L515 [58], a solid-state LiDAR camera. The camera is mounted on the robot head to provide egocentric vision. Previous studies have demonstrated that cameras with less accurate depth sensing, such as the RealSense D435 [59], can result in suboptimal performance for DP3 [17], [60]. It is important to note, however, that even the RealSense L515 does not produce perfectly accurate point clouds.

Height-Adjustable Cart. A major challenge in generalizing manipulation skills to real-world environments is the wide variation in scene conditions, particularly *the differing heights of tabletops*. To address this, we utilize a height-adjustable cart, eliminating the need for complex whole-body control. While this simplifies the manipulation process, we believe our approach will perform equally well once whole-body control techniques become more mature.

B. Data

Whole-Upper-Body Teleoperation. To teleoperate the robot's upper body, we employ the Apple Vision Pro (AVP, [61]), which provides precise tracking of the human hand, wrist, and head poses [62]. The robot uses Relaxed IK [63] to follow these poses accurately. We also stream the robot's vision back to the AVP. Different from [9], we incorporate the waist into our teleoperation pipeline, enabling a more flexible workspace.

Latency of Teleoperation. The use of a LiDAR sensor significantly occupies the bandwidth/CPU of the onboard computer, resulting in a teleoperation latency of approximately 0.5 seconds. We also try two LiDAR sensors (one additionally mounted on the wrist), which introduce extremely high latency and thus make the data collection infeasible.

Data for Learning. We collect trajectories of observation-action pairs during teleoperation, where observations consist of two parts: 1) visual data, such as point clouds and images, and 2) proprioceptive data, such as robot joint positions. Actions are represented by the target joint positions. We also tried using end-effector poses as proprioceptions/actions, finding no significant difference in performance.

C. Learning and Deployment

We train iDP3 on our collected human demonstrations. Notably, we do not rely on camera calibration or manual point cloud segmentation as mentioned before. Therefore, our iDP3 policy can be seamlessly transferred to new scenes without requiring additional efforts such as calibration/segmentation.

TABLE I: **Efficiency of iDP3 compared to baselines.** To improve the robustness of the baselines, we have added Random Crop and Color Jitter augmentation to all image-based methods during training. **All the methods are evaluated with more than 100 trials,** ensuring less randomness in real-world evaluation.

Baselines	DP	DP (❄R3M)	DP (✶R3M)	iDP3 (DP3 Encoder)	**iDP3**
1st-1	0/0	11/33	24/39	15/34	21/38
1st-2	7/34	10/28	27/36	12/27	19/30
3rd-1	7/36	18/38	26/38	15/32	19/34
3rd-2	10/36	23/39	22/34	16/34	16/37
Total	24/106	62/138	**99/147**	58/127	**75/139**

V. Experiments and Analysis

To evaluate the effectiveness of our system, our experiments will use the fundamental task of **Pick&Place** as the primary benchmark for our analysis.

A. Experiment Setup

Task Description. In this task, the robot grasps a lightweight cup and moves it aside. The challenge for humanoid robots with dexterous hands is that the cup is similar in size to the hands; thus, even small errors result in collisions or missed grasps. This task requires more precision than using parallel grippers, which can open wider to avoid collisions.

Task Setting. We train the Pick&Place task under four settings: {1st-1, 1st-2, 3rd-1, 3rd-2}. "1st" uses an egocentric view, and "3rd" uses a third-person view. The numbers behind represent the number of demonstrations used for training, with each demonstration consisting of 20 rounds of successful execution. The training dataset is kept small to highlight the differences between methods. The object position is randomly sampled in a 10cm×20cm region.

Evaluation Metric. We run three episodes for each method, each consisting of 1,000 action steps. In total, each method is evaluated with around 130 trials, ensuring a thorough evaluation of each method. We record both the number of successful grasps and the total number of grasp attempts. The successful grasp count reflects the accuracy of the policy. The total number of attempts serves as a measure of the policy's smoothness, since the jittering policies tend to hang around and have few attempts as we observe in experiments.

B. Effectiveness of iDP3

We compare iDP3 with several strong baselines, including: a) **DP**: Diffusion Policy [12] with a ResNet18 encoder; b) **DP (❄R3M)**: Diffusion Policy with a frozen R3M [64] encoder; c) **DP (✶R3M)**: Diffusion Policy with a finetuned R3M encoder; and d) **iDP3 (DP3 Encoder)**: iDP3 using the DP3 encoder [12]. All image-based methods use the same policy backbone as iDP3 and Random Crop and Color Jitter augmentations to improve robustness and generalization. The

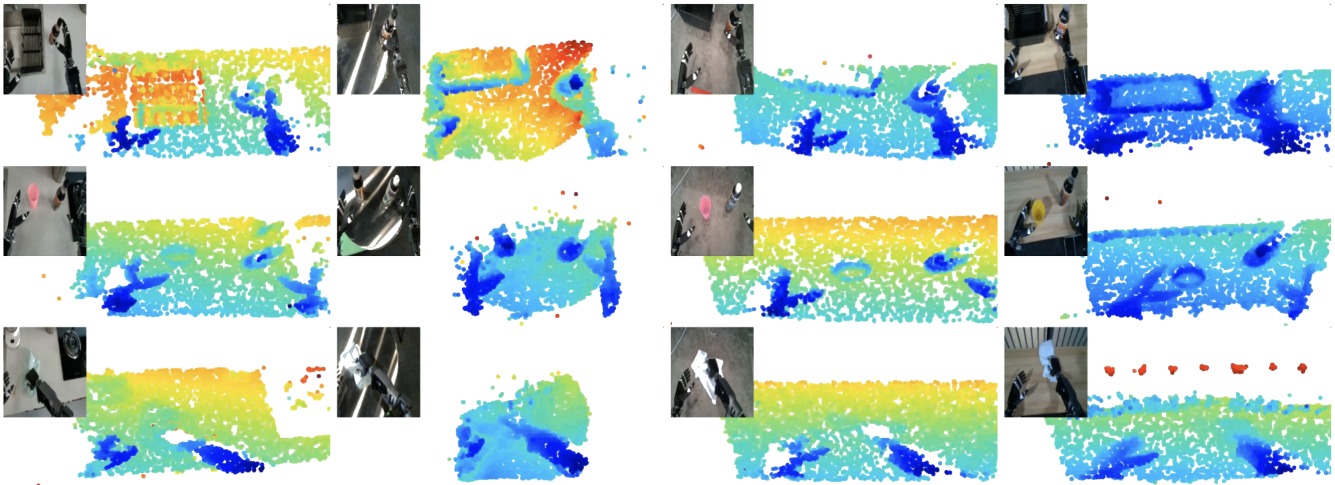

Fig. 4: **Visualization of egocentric 2D and 3D observations.** This figure highlights the complexity of diverse real-world scenes. Videos are available on our website.

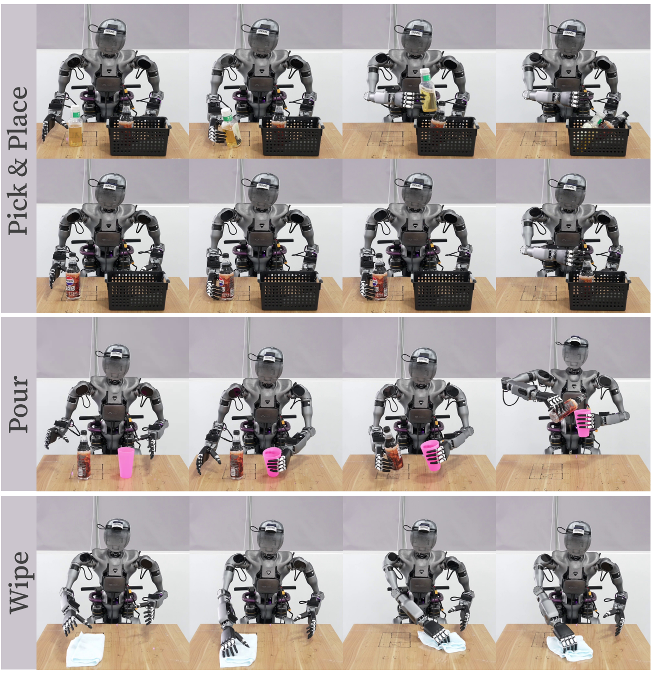

Fig. 5: **Trajectories of our three tasks in the training scene,** including Pick&Place, Pour, and Wipe. We carefully select daily tasks so that they are useful across scenes.

RGB image resolution is 224 × 224, resized from the raw image from the RealSense camera.

The results, presented in Table I, show that iDP3 significantly outperforms vanilla DP, DP with a frozen R3M encoder, and iDP3 with the DP3 encoder. However, we find that DP with a finetuned R3M is a particularly strong baseline, outperforming iDP3 in these settings. We hypothesize that this is because finetuning pre-trained models are often more effective compared to training-from-scratch [26], and there are currently no similar pre-trained 3D visual models for robotics.

Though DP+finetuned R3M is more effective in these settings, we find that image-based methods are overfitting to the specific scenario and object, failing to generalize to wild scenarios, as shown in Section VI.

TABLE II: **Ablation on iDP3.** The results demonstrate that removing certain key modifications from iDP3 significantly impacts the performance of DP3, leading to either failure in learning from human data or reduced accuracy. **All the methods are evaluated with more than 100 trials,** ensuring less randomness in real-world evaluation.

Visual Encoder	1st-1	1st-2	3rd-1	3rd-2	Total
Linear (DP3)	15/34	12/27	15/32	16/34	58/127
Conv	9/33	14/32	14/33	12/33	49/131
Linear+Pyramid	15/34	**20/31**	13/33	**18/36**	66/134
Conv+Pyramid (iDP3)	**21/38**	19/30	**19/34**	16/37	**75/139**

Number of Points	1st-1	1st-2	3rd-1	3rd-2	Total
1024 (DP3)	11/28	10/30	18/35	17/36	56/129
2048	17/35	13/28	17/32	**18/33**	65/128
4096 (iDP3)	21/38	**19/30**	**19/34**	16/37	**75/139**
8192	**24/35**	16/28	14/33	**18/36**	72/132

Prediction Horizon	1st-1	1st-2	3rd-1	3rd-2	Total
4 (DP3)	0/0	0/0	0/0	0/0	0/0
8	0/0	3/18	18/36	12/34	33/88
16 (iDP3)	**21/38**	19/30	**19/34**	**16/37**	**75/139**
32	9/34	**20/30**	14/33	12/33	55/130

Additionally, we believe there is still room for improvement in iDP3. Our current 3D visual observations are quite noisy due to the limitations of the sensing hardware. We expect that more accurate 3D observations could lead to optimal performance in 3D visuomotor policies, as demonstrated in simulation [17].

C. Ablations on iDP3

We conduct ablation studies on several modifications to DP3, including improved visual encoders, scaled visual input, and a longer prediction horizon. Our results, given in Table II, demonstrate that without these modifications DP3 either fails to learn effectively from human data or exhibits significantly reduced accuracy.

More specifically, we observe that 1) our improved visual encoder could both improve the smoothness and accuracy

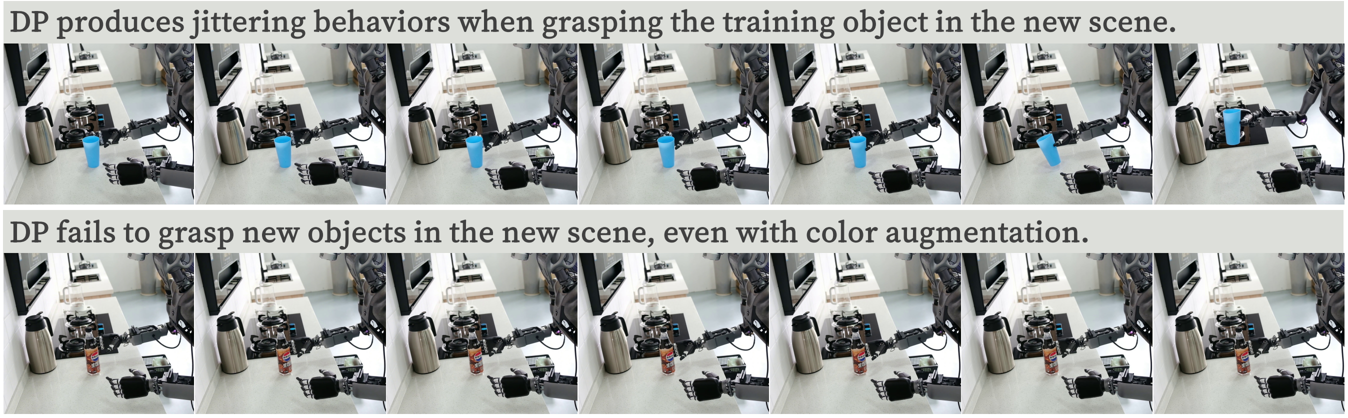
DP produces jittering behaviors when grasping the training object in the new scene.
DP fails to grasp new objects in the new scene, even with color augmentation.

Fig. 6: **Failure cases of image-based methods in new scenes.** Here DP corresponds to **DP (✶R3M)** in Table I, which is the strongest image-based baseline we have. We find that even added with color augmentation during training, image-based methods still struggle in the new scene/object.

TABLE III: **Capabilities of iDP3.** While iDP3 maintains similar efficiency to DP (✶R3M) (abbreviated as DP), it stands out with remarkable generalization capabilities, making it well-suited for real-world deployment. For evaluation in the new scene, we use the kitchen scene shown in Figure 6 and unseen objects are also included. We do not test Wipe in generalization settings since Wipe is achieved with high success rates for all methods.

Training	DP	**iDP3**	**New Object**	DP	**iDP3**	**New View**	DP	**iDP3**	**New Scene**	DP	**iDP3**
Pick&Place	**9/10**	**9/10**	Pick&Place	3/10	**9/10**	Pick&Place	2/10	**9/10**	Pick&Place	2/10	**9/10**
Pour	**9/10**	**9/10**	Pour	1/10	**9/10**	Pour	0/10	**9/10**	Pour	1/10	**9/10**
Wipe	**10/10**	**10/10**	Wipe	–	–	Wipe	–	–	Wipe	–	–

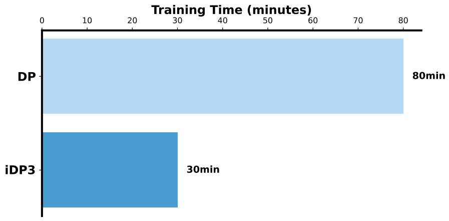

Fig. 7: **Training time.** Due to using 3D representations, iDP3 saves training time compared to Diffusion Policy (DP), even after we scale up the 3D vision input. This advantage becomes more evident when the number of demonstrations gets large.

of the policy; 2) scaled vision inputs are helpful, while the performance gets saturated in our tasks with more points; 3) an appropriate prediction horizon is critical, without which DP3 fails to learn from human demonstrations.

Additionally, Figure 7 presents the training time for iDP3, demonstrating a significant reduction compared to Diffusion Policy. This efficiency is maintained even when the number of point clouds increases to several times that of DP3 [17].

VI. Capabilities

In this section, we show more capabilities of iDP3 on humanoid robots. We also conduct more comparisons between iDP3 and DP (✶R3M) (abbreviated as DP in this section) and show that iDP3 is more applicable in the challenging and complex real world. Results are given in Table III.

Tasks. We select three tasks, **Pick&Place**, **Pour**, and **Wipe**, to demonstrate the capabilities of our system. We ensure that these tasks are common in daily life and could be useful for humans. For instance, Pour is frequently performed in restaurants, and Wipe in cleaning tables in households.

Data. We collect 10 demonstrations for each task. For Pick&Place task, each demonstration contains 10 trajectories of pick&place. In each demonstration, the object poses are randomized, limited in a region of 10cm×10cm. We do not collect data in a larger region, since we find that a larger task region simply requires more data [65]. Besides, collecting large-scale data is not feasible due to the usage of AVP.

Effectiveness. As shown in Table III, both iDP3 and DP achieve high success rates in the training environment with the training objects.

Property 1: View Invariance. Our egocentric 3D representations demonstrate impressive view invariance. As shown in Figure 8, iDP3 consistently grasps objects even under large view changes, while DP struggles to grasp even the training objects. DP shows occasional success only with minor view changes. Notably, unlike recent works [22], [45], [47], we did not incorporate specific designs for equivariance or invariance.

Property 2: Object Generalization. We evaluated new kinds of cups/bottles beside the training cup, as shown in Figure 9. While DP, due to the use of Color Jitter augmentation, can occasionally handle unseen objects, it does so with a low success rate. In contrast, iDP3 naturally handles a wide range of objects, thanks to its use of 3D representations.

Property 3: Scene Generalization. We further deploy our policy in various real-world scenarios, as shown in Figure 1. These scenes are nearby the lab and **none of the scenes**

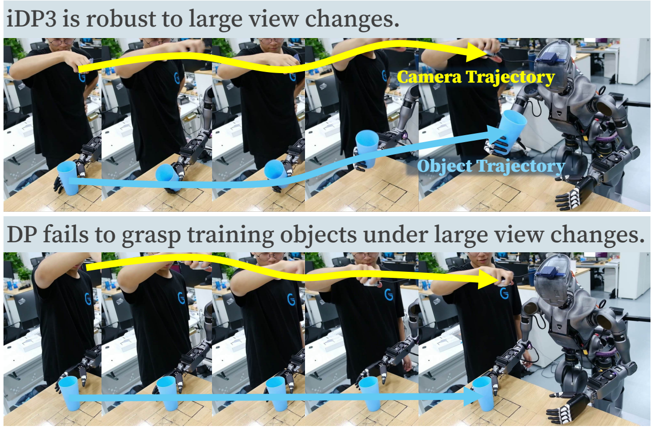
iDP3 is robust to large view changes.
Camera Trajectory
Object Trajectory
DP fails to grasp training objects under large view changes.

Fig. 8: **View invariance of iDP3.** We find that egocentric 3D representations are surprisingly view-invariant. Here DP corresponds to **DP (★R3M)** in Table I, which is the strongest image-based baseline we have.

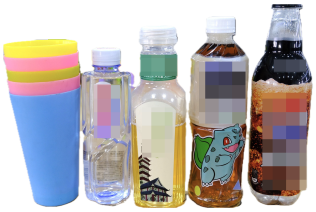

Fig. 9: **Objects used in Pick&Place and Pour.** We only use the cups as the training objects, while our method naturally handles other unseen bottles/cups.

are cherry-picked. The real world is far noisier and more complex than the controlled tabletop environments used in the lab, leading to reduced accuracy for image-based methods (Figure 6). Unlike DP, iDP3 demonstrates surprising robustness across all scenes. Additionally, we provide visualizations of both 2D and 3D observations in Figure 4.

VII. Conclusions and Limitations

Conclusions. This work presents an imitation learning system that enables a full-sized humanoid robot to generalize practical manipulation skills to diverse real-world environments, trained with data collected solely in the lab. The key is the Improved 3D Diffusion Policy (iDP3), a new 3D visuomotor policy for general-purpose robots. Through extensive experiments, we demonstrate the impressive generalization capabilities of iDP3 in the real world.

Limitations. 1) Teleoperation with AVP is easy to set up but tiring for human teleoperators, making data scaling infeasible. 2) The depth sensor produces noisy point clouds, limiting the performance of iDP3. 3) Collecting fine-grained manipulation skills, such as turning a screw, is time-consuming due to teleoperation with AVP. 4) We avoided using the robot's lower body, as maintaining balance is still challenging. In general, scaling up high-quality data is the main bottleneck. In the future, we hope to explore how to scale up the training of 3D visuomotor policies with more high-quality data.

Acknowledgments

We would like to thank Jie Gu, Bin Zhou, and Yusheng Cai from Fourier Intelligence for hardware support, Yuxiang Gao for help in teleoperation, Shuo Hu for help in the 3D print of the camera mount, Zeqian Bao, Renzhi Tao, and Jiayan Gu for helpful discussions. Besides, we would like to thank Chen Wang, Yunzhi Zhang, Zizhang Li, and Haoyu Xiong from Stanford University for their insightful discussions.

References

[1] Boston Dynamics, "Atlas," 2024, online. [Online]. Available: https://bostondynamics.com/atlas/

[2] Tesla, "Optimus," 2024, online. [Online]. Available: https://www.tesla.com/en_eu/AI

[3] Figure, "01," 2024, online. [Online]. Available: https://www.figure.ai/

[4] Unitree, "H1," 2024, online. [Online]. Available: https://www.unitree.com/h1

[5] Fourier Intelligence, "Gr1," 2024, online. [Online]. Available: https://www.fourierintelligence.com/gr1

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

[7] 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 *arXiv*, 2024.

[8] 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," in *arXiv*, 2024.

[9] 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.

[10] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *arXiv preprint arXiv:2407.03162*, 2024.

[11] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," *arXiv preprint arXiv:2408.11805*, 2024.

[12] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

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

[15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[16] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[17] 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.

[18] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *Arxiv*, 2024.

[19] M. Grotz, M. Shridhar, T. Asfour, and D. Fox, "Peract2: A perceiver actor framework for bimanual manipulation tasks," *arXiv preprint arXiv:2407.00278*, 2024.

[20] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 785–799.

[21] Y. Ze, G. Yan, Y.-H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L. E. Li, and X. Wang, "Gnfactor: Multi-task real robot learning with generalizable neural feature fields," in *Conference on Robot Learning*. PMLR, 2023, pp. 284–301.

[22] J. Yang, Z. ang Cao, C. Deng, R. Antonova, S. Song, and J. Bohg, "Equibot: Sim(3)-equivariant diffusion policy for generalizable and data efficient learning," *arXiv preprint arXiv:2407.01479*, 2024.

[23] Z. He, K. Lei, Y. Ze, K. Sreenath, Z. Li, and H. Xu, "Learning visual quadrupedal loco-manipulation from demonstrations," *arXiv preprint arXiv:2403.20328*, 2024.

[24] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[25] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[26] N. Hansen, Z. Yuan, Y. Ze, T. Mu, A. Rajeswaran, H. Su, H. Xu, and X. Wang, "On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline," *arXiv preprint arXiv:2212.05749*, 2022.

[27] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[28] Y. Ze, Y. Liu, R. Shi, J. Qin, Z. Yuan, J. Wang, and H. Xu, "H-index: Visual reinforcement learning with hand-informed representations for dexterous manipulation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.

[30] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.

[31] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[32] 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.

[33] H. Xiong, R. Mendonca, K. Shaw, and D. Pathak, "Adaptive mobile manipulation for articulated objects in the open world," *arXiv preprint arXiv:2401.14403*, 2024.

[34] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," in *RSS*, 2022.

[35] Y. Ze, N. Hansen, Y. Chen, M. Jain, and X. Wang, "Visual reinforcement learning with self-supervised 3d representations," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2890–2897, 2023.

[36] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[37] Z. Yuan, T. Wei, S. Cheng, G. Zhang, Y. Chen, and H. Xu, "Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning," *arXiv preprint arXiv:2407.15815*, 2024.

[38] Y. Jiang, C. Wang, R. Zhang, J. Wu, and L. Fei-Fei, "Transic: Sim-to-real policy transfer by learning from online correction," *arXiv preprint arXiv:2405.10315*, 2024.

[39] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 594–605.

[40] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," *arXiv preprint arXiv:2403.16967*, 2024.

[41] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," *arXiv preprint arXiv:2407.10353*, 2024.

[42] J. Wang, Y. Yuan, H. Che, H. Qi, Y. Ma, J. Malik, and X. Wang, "Lessons from learning to spin "pens"," *arXiv:2407.18902*, 2024.

[43] Y. Zhang, T. Liang, Z. Chen, Y. Ze, and H. Xu, "Catch it! learning to catch in flight with mobile dexterous hands," *arXiv preprint arXiv:2409.10319*, 2024.

[44] T. Lin, Z.-H. Yin, H. Qi, P. Abbeel, and J. Malik, "Twisting lids off with two hands," *arXiv:2403.02338*, 2024.

[45] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt, "Equivariant diffusion policy," *arXiv preprint arXiv:2407.01812*, 2024.

[46] Y. Wang, G. Yin, B. Huang, T. Kelestemur, J. Wang, and Y. Li, "GenDP: 3d semantic fields for category-level generalizable diffusion policy," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=7wMlwhCvjS

[47] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu, "View-invariant policy learning via zero-shot novel view synthesis," *arXiv preprint arXiv:2409.03685*, 2024.

[48] H. Etukuru, N. Naka, Z. Hu, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiullah, "General policies for zero-shot deployment in new environments," *arXiv*, 2024.

[49] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," *arXiv preprint arXiv:2408.14472*, 2024.

[50] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik, "Humanoid locomotion as next token prediction," *arXiv preprint arXiv:2402.19469*, 2024.

[51] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[52] 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.

[53] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[54] 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 *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2023.

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[56] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[57] Inspire Robots, "Dexterous hands," 2024, online. [Online]. Available: http://www.inspire-robots.store/collections/the-dexterous-hands

[58] Intel RealSense, "Lidar camera l515," 2024, online. [Online]. Available: https://www.intelrealsense.com/lidar-camera-l515/

[59] ——, "Depth camera d435," 2024, online. [Online]. Available: https://www.intelrealsense.com/depth-camera-d435/

[60] C. Wang, H. Fang, H.-S. Fang, and C. Lu, "Rise: 3d perception makes real-world robot imitation simple and effective," *arXiv preprint arXiv:2404.12281*, 2024.

[61] Apple, "Apple vision pro," 2024, online. [Online]. Available: https://www.apple.com/apple-vision-pro/

[62] Y. Park and P. Agrawal, "Using apple vision pro to train and control robots," 2024, online. [Online]. Available: https://github.com/Improbable-AI/VisionProTeleop

[63] D. Rakita, B. Mutlu, and M. Gleicher, "Relaxedik: Real-time synthesis of accurate and feasible robot arm motion." in *Robotics: Science and Systems*, vol. 14. Pittsburgh, PA, 2018, pp. 26–30.

[64] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[65] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "ALOHA unleashed: A simple recipe for robot dexterity," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=gvdXE7ikHI