2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

FP3: A 3D Foundation Policy for Robotic Manipulation

Rujia Yang*[1], Geng Chen*[4], Chuan Wen†[3], Yang Gao†[1,2]

[1]Tsinghua University [2]Shanghai Qi Zhi Institute
[3]Shanghai Jiao Tong University [4]UC San Diego

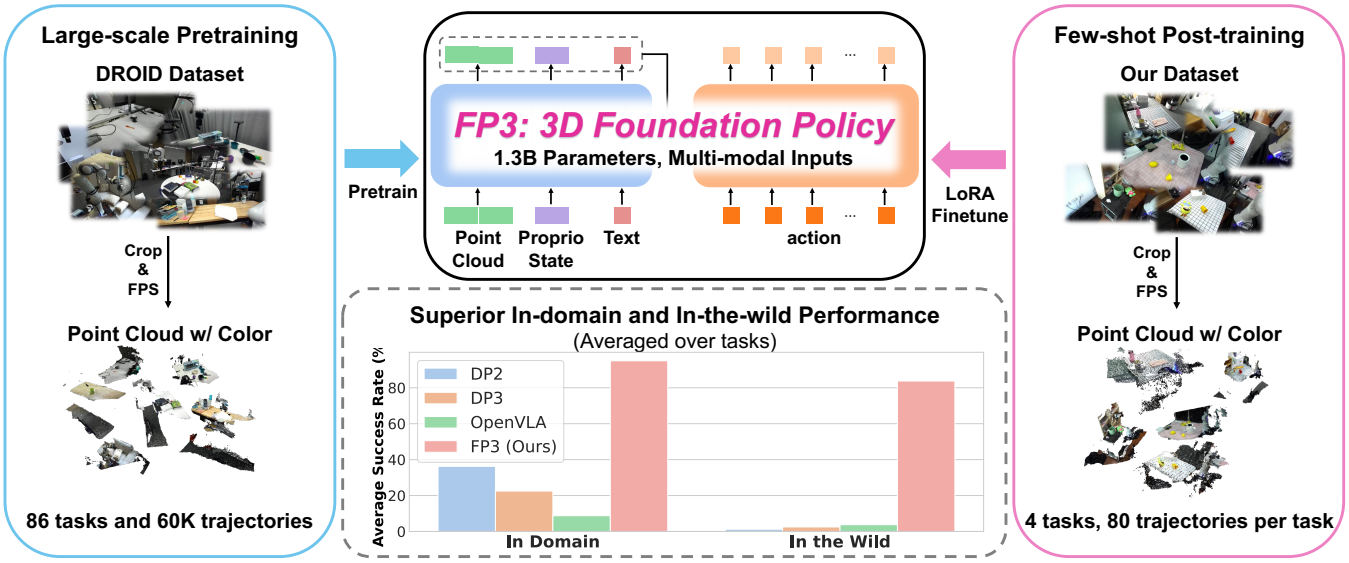

Fig. 1: **Overview of 3D Foundation Policy (FP3)**, a 1.3B 3D point cloud-based language-visuomotor policy pre-trained on 60k episodes from the DROID dataset [1]. FP3 supports data-efficient fine-tuning for downstream tasks, while demonstrating superior generalizability to unseen environments and novel objects.

***Abstract*— Following its success in natural language processing and computer vision, foundation models that are pre-trained on large-scale multi-task datasets have also shown great potential in robotics. However, most existing robot foundation models rely solely on 2D image observations, ignoring 3D geometric information, which is essential for robots to perceive and reason about the 3D world. In this paper, we introduce FP3, a large-scale 3D foundation policy model for robotic manipulation. FP3 builds on a scalable diffusion transformer architecture and is pre-trained on 60k trajectories with point cloud observations. With the model design and diverse pre-training data, FP3 can be efficiently fine-tuned for downstream tasks while exhibiting strong generalization capabilities. Experiments on real robots demonstrate that with only 80 demonstrations, FP3 is able to learn a new task with over 90% success rates in novel environments with unseen objects, significantly surpassing existing robot foundation models.**

I. Introduction

Learning-based policies have shown great effectiveness in robotic manipulation [2], [3], [4], [5], [6], [7]. However, these learned policies often show limited or even zero generalization capability to unseen scenarios, new objects, and distractors [8]. Additionally, most current methods are

*Equal contribution.
†Equal advising. Corresponding authors.

trained on single or few tasks[4], [5], requiring a relatively large amount of expert demonstrations (usually about 200 episodes) to learn a new task. In contrast, natural language processing (NLP) and computer vision (CV) have achieved remarkable success in developing foundation models that are trained on large-scale data and diverse tasks, enabling them to generalize to arbitrary scenarios in the wild. Therefore, building a similar foundation model in robotic manipulation that can generalize to novel objects, scenes, and tasks becomes a promising topic [6], [9], [7].

Towards this goal of policy foundation models, there have been some initial attempts at vision-language-action (VLA) models [3], [6], [7], [10], [11], [12], which build on vision-language models (VLM) trained on internet-scale vision and language data to inherit commonsense knowledge and fine-tune VLMs on large-scale robotics datasets [13], [1]. Despite significant progress, their generalizability remains limited when confronted with novel tasks, objects, scenes, and camera views, etc.

One potential limitation of current policy foundation models is their exclusive reliance on 2D image observations, lacking 3D observation inputs. However, 3D geometric information is significant for perceiving 3D environments and reasoning about spatial relationships [14], [15], [16], [17]. Some works have shown that 3D representations can improve the sample efficiency and generalizability of robotic manipulation policies [18], [16], [19], [15]. Among all the 3D representations such as RGB-D images, point clouds, voxels, and 3D Gaussians [20], point clouds are found to be the most effective [16], [14].

In this work, we introduce 3D Foundation Policy (FP3), the first 3D point cloud-based language-visuomotor policy foundation model for robotic manipulation that exhibits strong generalizability and sample efficiency. To extract rich semantic and geometric representation from 3D point cloud observation, FP3 adopts a pre-trained large-scale point cloud encoder Uni3D [21]. We further leverage an encoder-decoder Diffusion Transformer (DiT) architecture to integrate the point cloud representations, language embedding, and proprioception for denoising the actions.

With the proposed policy architecture, we employ a pre-training&post-training recipe for FP3, mirroring the common practice in large language models (LLMs) [22], [23] where the model is pre-trained on large-scale diverse corpus and fine-tuned on curated task-specific data to adapt for downstream tasks. We first pre-train FP3 on the large-scale robotic manipulation dataset DROID [1] which contains 76k demonstration trajectories or 350h of interaction data from 564 scenes and 86 tasks. We then collect a small amount of high-quality teleoperation data for several tasks and fine-tune FP3. The result indicates that our model can efficiently master a new task with only 80 post-training trajectories and is capable of zero-shot generalization to both novel objects and environments with about 90% success rate. In contrast, strong baselines like DP3 [16] and OpenVLA [6] almost completely fail in this setting. Due to the advantages of 3D representation, FP3 is also robust to background variations, lighting conditions, camera angles, and distractors. Additionally, we demonstrate FP3's ability to handle simple tasks in unseen environment *zero-shot* without any fine-tuning. Finally, we conduct ablation studies to demonstrate that the 3D representation, data scaling, and model scaling all contribute to the model's superior performance.

We summarize our main contributions as follows:

1) We propose a novel diffusion-based 3D robot policy architecture, FP3.
2) We pre-train FP3 on large-scale robotic manipulation data with 3D observation, establishing a 1B-parameter 3D policy foundation model.
3) We collect data on several new tasks and demonstrate the efficient and generalizable fine-tuning of FP3, achieving around 60% in-domain and 80% in-the-wild performance improvement on average over strong baselines with only 2-hour and single-GPU finetuning.

II. Related Work

A. Foundation Models in Robotics

Similar to the cases in natural language processing and computer vision, foundation models have already been widely used in multiple aspects of robotics, including representation learning [24], [25], [26], [27], high-level task planning [28], [29], training-free robotic manipulation [30], [31], [32], etc. In this work, we concentrate on policy foundation models in robotics, which also often refer to multi-task "generalist" robot policies [2], [3], [6], [33], [9], [7], [11], [12] trained on large-scale robot datasets [13], [34], [1], [35]. A significant subset of policy foundation models consists of the vision-language-action (VLA) models like RT-2 [3], OpenVLA [6] and π_0 [7], [10], which fine-tune pre-trained large vision-language models to predict actions by treating discretized actions [3], [6], [10] as language tokens or adding a diffusion-based action expert [7], [11], [12]. Our work is most similar to RDT [9] in architecture, which both scale up a diffusion transformer to predict actions, yet differs in key areas like conditioning blocks and embodiment.

In addition to architecture, a key difference between these approaches and FP3 is the observation modality. Unlike these works, which all take 2D images as input, our work utilizes 3D point clouds to enhance the perception of 3D geometric information and reasoning about spatial relationships.

B. Robotic Manipulation with 3D representations

Compared to 2D images, 3D representations such as RGB-D images, point clouds, and voxels contain richer geometric information and have thus been widely used in robotic manipulation [18], [16], [36], [37]. Kite [38] directly leverages the RGB-D observation for semantic manipulation. Other works [39], [40], [16], [41], [42] reconstruct point clouds from RGB-D images and process them using a point cloud encoder for manipulation. Voxelizing the point clouds for perception is also a viable solution [43], [18], [44]. Another set of works[45], [46], [19], [15], [47] lifts 2D image features to 3D space to benefit from both semantic and geometric information. There have also been attempts

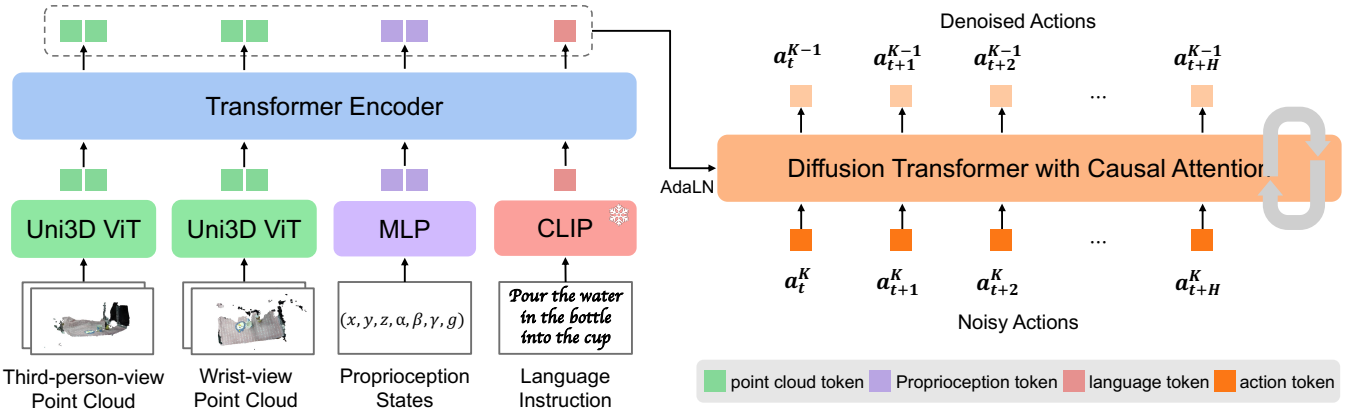

Fig. 2: **FP3 architecture.** Each camera view's point cloud observation $\mathbf{P}_t^i$ (with history length of two) is encoded with a Uni3D ViT-L [21] encoder. The language instruction ℓ_t is embedded with a frozen CLIP [66] model and a text embeddings E5 [67]. The Transformer encoder fuses multi-modal input embeddings to latent tokens, while the Transformer decoder takes in the noise actions and leverages adaLN [68], [69], [70] blocks to integrate the latent tokens generated by the encoder, predicting denoised action chunks.

in combining implicit or explicit 3D reconstruction (NeRF [48], 3D gaussians [20]) with robotic manipulation [49], [50], [51], [52], [53]. In this work, we choose the point cloud as the 3D representation as it is found to be more effective than other representations in DP3 [16].

C. Diffusion models in Robotics

Diffusion models have achieved great success in image generation [54], [55], [56] and video generation [57], [58] by modeling complex high-dimensional distributions through progressive denoising. Due to their remarkable expressiveness, diffusion models have also been applied across various fields in robotics, including reinforcement learning [59], [60], [61], imitation learning [4], [16], [62], [9], [33], [7], and motion planning [63], [64], [65]. Our work focuses on "Diffusion Policies" [4], [16], [9], which refers to methods thatemploy conditional diffusion models as visuomotor policy models for imitation learning. A key distinction between our work and previous diffusion-based policies is that our FP3 model leverages 3D point cloud representations to improve data efficiency and generalizability.

III. Method

We now introduce the 3D Foundation Policy (FP3) model for generalist robotic manipulation, achieving high data efficiency and generalization capability. FP3 is a 1.3B encoder-decoder transformer network following a two-stage pre-training and post-training recipe. We first provide the detailed architecture and key design decisions of FP3 in Section III-A. Then we describe the pre-training and post-training procedures in Section III-B and Section III-C, respectively.

A. FP3 model

At its core, FP3 is a diffusion-based policy model similar to [4], [9]. It takes the 3D point cloud observation, language, and robot proprioceptive state as input and predicts action chunks of future actions. Formally, we formalize the problem of language-conditioned visuomotor control as modeling the distribution $p(A_t|o_t)$, where $\mathbf{o}_t = [\mathbf{P}_t^1, ..., \mathbf{P}_t^n, \ell_t, \mathbf{q}_t]$ is the observation at time t including point cloud observation $\mathbf{P}_t^i$ from the i^{th} camera (including historical observation), language instruction ℓ_t and proprioceptive information $\mathbf{q}_t$ and $A_t = [a_t, a_{t+1}, ..., a_{t+H-1}]$ denotes the predicted action chunk. We train a denoising diffusion probabilistic model (DDPM) [54] to approximate the conditional distribution and use the denoising diffusion implicit model (DDIM) [55] method to accelerate inference. Next we describe the detailed structure of FP3 model, including the encoding of multi-modal inputs and the transformer-based encoder-decoder architecture.

Encoding of multi-modal inputs. To process the multi-modal input, we encode the input signals into a unified token space with the same dimensions as follows:

- **Point cloud observations** contain rich semantic and geometric information and are found to be more suitable for policy learning compared to other 3D representations [16]. Therefore, we consider using point cloud as the 3D representation in FP3. Current point cloud-based robot policies [16], [41], [47] typically use sparse point cloud and small networks such as PointNet++ [71] and PointNeXt [72] to encode the points into embeddings. However, pre-trained large-scale foundation vision encoders have demonstrated a performance advantage over small encoders in image-based policies [73], [74]. Consequently, we increase the number of input points to 8000 for each view and employ a 300M-parameter point cloud encoder Uni3D ViT [21] that is pre-trained to align the 3D point cloud features with the image-text aligned features to obtain the point cloud embeddings. For the third-person-view and the wrist-view point clouds, we use separate encoders since their point distributions might be greatly different. Following [74], we choose to fine-tune the weights of Uni3D ViTs during policy training.
- **Language instructions** are both encoded with a CLIP

[66] model which aligns with Uni3D and a strong text embeddings E5 [67]. The outputs from both encoders are concatenated to create a single language embedding. Both weights are fixed during training.

- **Low-dimensional inputs** including robot proprioceptive state and noise levels are processed with two-layer MLPs, respectively.

Encoder-decoder structure. Since Diffusion Transformers have shown great scalability in image generation [68], [75] and policy learning [76], [77], we adopt the transformer architecture and scale it up for FP3. To better fuse the point cloud, language, and proprioceptive state embeddings, we utilize a Transformer Encoder-Decoder architecture similar to [5], [78], [62], [77]. Specifically, FP3 first feeds all the embeddings into a transformer encoder, producing a sequence of informative latent tokens.

The diffusion denoiser of FP3 is a Transformer decoder that denoises the action chunks from noise with temporal causal masking following [76]. To infuse the multi-modal latent tokens into the denoiser, FP3 adopts the adaptive LayerNorm (adaLN) module for conditioning, which is essential for implementing diffusion training in image generation [68], [75] and policy learning [76], [77].

B. Pre-training

Pre-training data. To build a 3D policy foundation model, we need to train our model on large-scale 3D robotic manipulation datasets. However, most existing large-scale robot datasets such as the Open X-Embodiment dataset (OXE, [13]) are 2D-only. In this work, we choose the DROID dataset [1], which includes 86 tasks and 76k demonstrations, and use the 60k demonstrations from DROID that have depth observations to pre-train FP3.

Data pre-processing. Following DROID, we take the observations of two third-view cameras and one wrist-view camera in FP3. We use the RGB image and the depth map to recover the 3D point cloud for each camera and transform the two point clouds to the same world frame. As we only care about the operated object, we cropped the points outside a 1-meter box to remove redundant points. Further, we downsample each point cloud by farthest point sampling (FPS, [79]) to 8000 points to facilitate model training while retaining sufficient information. We preserve the color channels of each point to enable further experiments conditioned on colors.

Pre-training details. Following prior works [6], [74] which found that freezing pre-trained vision encoders may harm the policy performance, we fine-tune the Uni3D ViT encoder during pre-training. We use the absolute end-effector position as our action space. To enhance the model's robustness, we apply an independent 10% masking probability to each input component and randomly drop 0% to 80% of points during pre-training.

C. Post-training

After obtaining the pre-trained base model, we further employ a post-training process using a small amount of high-quality data to adapt the model to certain tasks, which aligns with most modern LLM practice [22], [23]. Different from the fine-tuning settings adopted in most existing robot foundation models in which they focus on either fine-tuning the model to adapt to new robot setups [6], [33] or learning new tasks in a fixed environment [9], [7], our goal is to fine-tune our model to *solve specific tasks on any object, in any environment*.

To achieve this goal, we further collect data for each downstream task in our robot setup. Taking the lesson from [74], we aim to enhance the diversity of environments and objects rather than merely increasing the number of demonstrations in the same scenario. Specifically, for each task, we collect 10 teleoperation demonstrations in each of 8 environments using 8 unique objects, i.e., 80 demonstrations in total. We then fine-tune the base model on this data using the parameter-efficient fine-tuning strategy LoRA [80]. Thanks to the effective initialization from pre-training, this small amount of fine-tuning data enables zero-shot deployment to novel environments and objects. We will discuss the tasks and results in detail in Section IV.

IV. Experiments

We conduct experiments on real robots for four downstream tasks to investigate the following questions: (1) Can FP3 be efficiently fine-tuned for new tasks? (2) How well does the fine-tuned FP3 generalize to unseen objects and scenes compared to the existing imitation learning (foundation) policies? (3) How robust is FP3 to the environment perturbations, such as lighting, camera views, distractors, *etc.*? (4) Can FP3 correctly execute the corresponding tasks following the language instruction? (5) How does FP3 perform zero-shot in new environments without fine-tuning?

A. Experimental setups

Simulation benchmarks. For simulation experiments, we choose a widely used benchmark ManiSkill[81] and select five tasks—`Peg Insertion`, `Stack Cube`, `Pick Cube`, `Push Cube`, and `Plug Charger`. We evaluate all the models in two settings, using 100 and 1000 demonstrations per task, respectively.

Real-world experiments. We build a real robot setup similar to DROID with a Franka robot arm and a Robotiq gripper to evaluate four downstream tasks as illustrated in Figure 3.

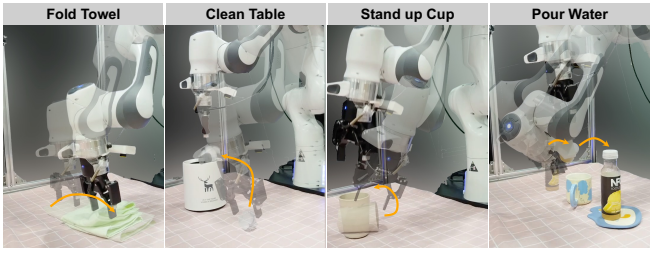

Fig. 3: **Real world task illustrations.** We evaluate our model on four downstream real-world tasks: `Fold Towel`, `Clean Table`, `Stand up Cup`, and `Pour Water`.

TABLE I: **Simulation post-training Evaluation.** We fine-tune FP3 and baseline methods on demonstrations from 5 tasks in Maniskill [81] environments. The table shows the success rate of all methods with different numbers of demos. Each result is the average success rate over 100 trials conducted with a different random seed.

	Peg Insertion		Stack Cube		Pick Cube		Push Cube		Plug Charger		Average	
	1k demos	100 demos	1k demos	100 demos	1k demos	100 demos	1k demos	100 demos	1k demos	100 demos	1k demos	100 demos
DP3	0.00	0.00	0.53	0.42	0.72	0.38	0.99	1.00	0.00	0.00	44.8	36.0
RDT	**0.05**	0.00	0.76	0.26	**0.78**	0.34	1.00	0.80	0.00	0.00	51.8	28.0
π_0	0.00	0.00	0.70	0.56	0.59	0.33	1.00	0.98	0.00	0.00	45.8	37.4
π_0-FAST-DROID	0.00	0.00	0.87	0.78	**0.85**	0.40	1.00	1.00	0.00	0.00	**54.5**	43.6
FP3 (**ours**)	0.04	**0.00**	**0.92**	**0.84**	0.68	**0.48**	**1.00**	**1.00**	**0.00**	**0.00**	52.8	**46.4**

Baselines. To comprehensively evaluate FP3, we carefully select three baselines for real-world experiments: Diffusion Policy (DP) [4]: a classic diffusion-based imitation learning policy with 2D image observation, DP3 [16]: an alternative version of DP which changes the 2D image observation to 3D point cloud, and OpenVLA [6]: a commonly-used image-based VLA model. We further add three baseline VLA models RDT [9], π_0 [7] and π_0-FAST-DROID [10] in simulation.

B. Efficient and generalizable fine-tuning for new tasks

1) Simulation experiments: We first evaluate FP3's capability to learn new skills efficiently in five simulation tasks. For each task, we use 100 and 1000 demonstrations to test the post-training performance in different downstream task data budgets. For DP3, we train the policy from scratch, while for RDT, π_0, and π_0-FAST-DROID, we follow the pre-training and post-training recipe to fine-tune the base models for each task with pre-trained weights.

Results in Table I show that in the 1000 demos setting, FP3 matches the frontier VLA models, and in the 100 demos setting, FP3 outperforms all comparing baselines. Notably, though π_0-FAST-DROID is pretrained on 1M demos before DROID fine-tuning, FP3 uses far less data yet fine-tunes more efficiently thanks to the 3D representations.

2) Real-world experiments: To comprehensively assess FP3's efficiency and generalizability, We further fine-tune FP3 in four real-world tasks. We collect only 10 demonstrations per environment-object pair for 8 pairs to obtain 80 demonstrations in total for each task. We not only evaluate all the policies in four in-domain environments with seen objects, but also deploy them zero-shot in four out-of-domain environments with unseen objects, which is a huge challenge for the model's generalizability.

In-domain Performance. Results in Table II show that in in-domain experiments, with only 10 demonstrations per scene, DP and DP3 can somewhat handle easier tasks, even though the success rate is below 50% in most cases; however, they almost completely fail in the more difficult task `Pour Water`. And OpenVLA struggles to perform any task, possibly due to the lack of action chunking (note that the seemingly poor performance of OpenVLA has also been reported by other works [9], [7]). In contrast, thanks to pre-training and 3D representation, FP3 efficiently learns all tasks with a success rate exceeding 90%.

In-the-Wild Performance. We further move the robot arm to novel environments and evaluate the policies with unseen objects. In this challenging setting, we observe that all baseline policies without pre-training, often fail to recognize the target objects, resulting in near-zero performance. In contrast, FP3 and consistently performs well with an average success rate of over 80%, devastating all the baselines. We attribute the superior performance to our extensive pre-training with diverse data, which enhances policy robustness. Furthermore, the geometric information in point cloud observation is also crucial for cross-domain generalization.

C. More experiments on generalization

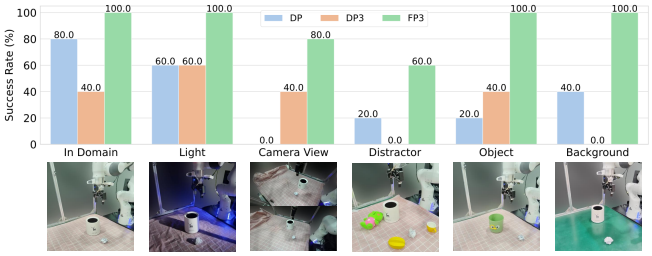

Fig. 4: **Generalization evaluation.** We evaluate FP3 and baseline policies on a diverse set of tasks, covering different axes of generalization.

Next we conduct more comprehensive experiments on FP3's generalizability to different environments and robot setups using the `Clean Table` task. Figure 4 demonstrates the results and visualizations.

Generalization to different object appearances, backgrounds, and lighting conditions. Conventional image-based policy networks are sensitive to visual variations. Therefore, we systematically evaluate the policies in different visual shifts. The results indicate that the DP's performance dropped hugely. In contrast, DP3 was more stable with lighting and color changes but was limited by in-domain performance. With the benefits of pre-training initialization and 3D geometry understanding, FP3 surpasses the baselines.

Generalization to new camera views. We also adjust the camera view by approximately 30 degrees from the training data to evaluate the camera view robustness of the policies. Once again, DP completely fails in this scenario, and DP3 is constrained by its in-domain performance, while FP3 maintains its high performance.

Generalization to distractors. We also try putting random distractors around the target object to assess the robustness

TABLE II: **Real-world post-training Evaluation.** We fine-tune FP3 and baselines on 80 real-world demonstrations from 8 environments and evaluate them on four in-domain environments with seen objects and four in-the-wild environments with unseen objects, conducting 5 trials for each. FP3 significantly outperforms other policies both in domain and in the wild.

	Fold Towel		Clean Table		Stand up Cup		Pour Water		Average	
	In-domain	In-the-wild	In-domain	In-the-wild	In-domain	In-the-wild	In-domain	In-the-wild	In-domain	In-the-wild
DP	20	0	75	5	45	0	5	0	36.25	1.25
DP3	20	0	25	10	45	0	0	0	22.50	2.50
OpenVLA	5	0	15	5	15	10	0	0	7.50	3.75
FP3-Scratch	35	5	35	0	15	0	35	0	30.00	1.25
FP3 (**ours**)	**90**	**85**	**100**	**95**	**95**	**75**	**95**	**75**	**95.00**	**82.50**

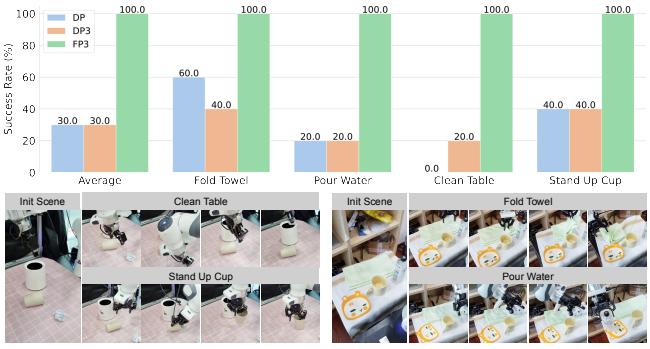

Fig. 5: **Instruction following evaluation.** We evaluate FP3 and baseline policies in the same initial state with different language instructions. FP3 can perfectly follow the instructions to execute the correct tasks rather than simply memorize the training distribution.

of these policies. In such settings, we find that the baseline policies may attempt to grasp interfering objects while FP3 remains most stable.

D. Instruction following

Since FP3 is a language-conditioned visuomotor policy, it's also important to evaluate its capability to execute tasks following the language command. Hence we fine-tune FP3 and the baselines in a multi-task setting with language. Figure 5 demonstrates that FP3 can execute tasks according to different language commands within the same starting context, while the baseline methods either fail to complete the task or are disrupted by the target objects of other tasks.

TABLE III: **Ablation study.** We achieve the best performance when using 3D point cloud input, a larger model, and larger-scale pre-training data.

	In-domain	In-the-wild
FP3-Scratch	35	0
FP3-Base-Image	90	55
FP3-Base	95	90
FP3-Base-30k	95	90
FP3 (**ours**)	**100**	**95**

E. Zero-shot Performance

We further evaluate FP3 against π_0-FAST-DROID *zero-shot* in an unseen environment by simply prompting it in

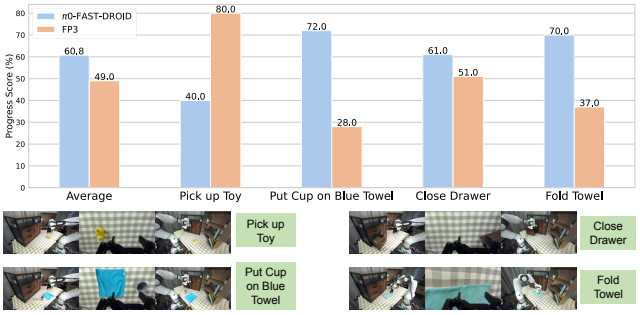

Fig. 6: **Zero-shot evaluation.** We evaluate FP3 and π_0-FAST-DROID *zero-shot* in a novel environment. FP3 shows comparable results to the SOTA VLA model.

natural language. We perform 10 rollouts for each model on four tabletop tasks and report a human tester-assigned score of task progress in Figure 6. Although still far from perfect, FP3 can solve simple tasks out of the box or at least make meaningful attempts, achieving results comparable to π_0-FAST. It's worth mentioning that such zero-shot task execution is quite challenging and is not demonstrated in the DROID [1], OpenVLA [6], or even π_0 paper [7].

F. Ablations

We finally do ablation studies on the observation choice, model size, and pre-training data size. We consider the following variants of FP3: *FP3-Base* reduces the model size to ViT-Base. *FP3-Base-30k* further reduces the pre-training data from 60k to 30k demonstrations. *FP3-Base-Image* converts the point cloud observations to images. Table III presents the results of each variant on the `Clean Table` task. Results show that the 3D modality, model size and data scale all contribute to the performance.

V. Conclusion

In this work, we present the 3D Foundation Policy (FP3), a large-scale Diffusion Transformer-based policy with 3D point cloud input. We pre-train FP3 on 60k episodes of robotic manipulation data and subsequently fine-tune it for downstream tasks. Through extensive experiments, we demonstrate that FP3 serves as an outstanding policy initialization for data-efficient and generalizable fine-tuning for new tasks. With only 80 demonstrations, FP3 can learn a new task with over 90% success rates in novel environments with unseen objects, significantly outperforming existing robot

policies. We hope that our work will pave the way for more exciting advancements in robot foundation models utilizing 3D representations.

VI. ACKNOWLEDGEMENT

This work is supported by Shanghai Qi Zhi Institute & Spirit AI Innovation Program and the Tsinghua University Dushi Program.

REFERENCES

[1] 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," CoRR, 2024.

[2] 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.

[3] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in Conference on Robot Learning. PMLR, 2023, pp. 2165–2183.

[4] 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.

[5] 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.

[6] 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.

[7] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al., "π_0: A vision-language-action flow model for general robot control," arXiv preprint arXiv:2410.24164, 2024.

[8] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 3153–3160.

[9] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," arXiv preprint arXiv:2410.07864, 2024.

[10] 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.

[11] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, et al., "Gemini robotics: Bringing ai into the physical world," arXiv preprint arXiv:2503.20020, 2025.

[12] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al., "Gr00t n1: An open foundation model for generalist humanoid robots," arXiv preprint arXiv:2503.14734, 2025.

[13] 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.

[14] H. Zhu, Y. Wang, D. Huang, W. Ye, W. Ouyang, and T. He, "Point cloud matters: Rethinking the impact of different observation spaces on robot learning," arXiv preprint arXiv:2402.02500, 2024.

[15] T. Zhang, Y. Hu, H. Cui, H. Zhao, and Y. Gao, "A universal semantic-geometric representation for robotic manipulation," in Conference on Robot Learning. PMLR, 2023, pp. 3342–3363.

[16] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation, 2024.

[17] C. Wen, D. Jayaraman, and Y. Gao, "Can transformers capture spatial relations between objects?" in The Twelfth International Conference on Learning Representations, 2023.

[18] 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.

[19] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," in First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024.

[20] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." ACM Trans. Graph., vol. 42, no. 4, pp. 139–1, 2023.

[21] J. Zhou, J. Wang, B. Ma, Y.-S. Liu, T. Huang, and X. Wang, "Uni3d: Exploring unified 3d representation at scale," in The Twelfth International Conference on Learning Representations.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[23] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., "Gemini: a family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.

[24] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in 6th Annual Conference on Robot Learning.

[25] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," in The Eleventh International Conference on Learning Representations.

[26] C. Wen, X. Lin, J. I. R. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point Trajectory Modeling for Policy Learning," in Proceedings of Robotics: Science and Systems, Delft, Netherlands, July 2024.

[27] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," in 8th Annual Conference on Robot Learning, 2024.

[28] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, et al., "Do as i can, not as i say: Grounding language in robotic affordances," in Conference on robot learning. PMLR, 2023, pp. 287–318.

[29] Y. Hu, F. Lin, T. Zhang, L. Yi, and Y. Gao, "Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning," in First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024.

[30] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in Conference on Robot Learning. PMLR, 2023, pp. 540–562.

[31] H. Huang, F. Lin, Y. Hu, S. Wang, and Y. Gao, "Copa: General robotic manipulation through spatial constraints of parts with foundation models," in First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024.

[32] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," in 8th Annual Conference on Robot Learning.

[33] 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.

[34] 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.

[35] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, "Rh20t: A robotic dataset for learning diverse skills in one-shot," in RSS 2023 Workshop on Learning for Task and Motion Planning, 2023.

[36] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," in Conference on Robot Learning. PMLR, 2023, pp. 694–710.

[37] P. Li, Y. Chen, H. Wu, X. Ma, X. Wu, Y. Huang, L. Wang, T. Kong, and T. Tan, "Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models," arXiv preprint arXiv:2506.07961, 2025.

[38] P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg, "Kite: Keypoint-conditioned policies for semantic manipulation," in Conference on Robot Learning. PMLR, 2023, pp. 1006–1021.

[39] S. Chen, R. G. Pinel, C. Schmid, and I. Laptev, "Polarnet: 3d point clouds for language-guided robotic manipulation," in Conference on Robot Learning. PMLR, 2023, pp. 1761–1781.

[40] W. Yuan, A. Murali, A. Mousavian, and D. Fox, "M2t2: Multi-task masked transformer for object-centric pick and place," in 7th Annual Conference on Robot Learning.

[41] C. Wang, H. Fang, H.-S. Fang, and C. Lu, "Rise: 3d perception makes real-world robot imitation simple and effective," in ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation.

[42] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," arXiv preprint arXiv:2503.07511, 2025.

[43] S. James, K. Wada, T. Laidlow, and A. J. Davison, "Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 739–13 748.

[44] H. Huang, O. Howell, X. Zhu, D. Wang, R. Walters, and R. Platt, "Fourier transporter: Bi-equivariant robotic manipulation in 3d," arXiv preprint arXiv:2401.12046, 2024.

[45] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," in 7th Annual Conference on Robot Learning, 2023.

[46] Z. Xian, N. Gkanatsios, T. Gervet, T.-W. Ke, and K. Fragkiadaki, "Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation," in 7th Annual Conference on Robot Learning, 2023.

[47] T. Zhang, Y. Hu, J. You, and Y. Gao, "Leveraging locality to boost sample efficiency in robotic manipulation," arXiv preprint arXiv:2406.10615, 2024.

[48] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[49] 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.

[50] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-nerf: Using a neural radiance field to grasp transparent objects," in 5th Annual Conference on Robot Learning.

[51] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 1757–1763.

[52] Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu, et al., "Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping," arXiv preprint arXiv:2403.09637, 2024.

[53] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation," in European Conference on Computer Vision. Springer, 2024, pp. 349–366.

[54] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[55] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in International Conference on Learning Representations.

[56] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695.

[57] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," Advances in Neural Information Processing Systems, vol. 35, pp. 8633–8646, 2022.

[58] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," arXiv preprint arXiv:2311.15127, 2023.

[59] Z. Wang, J. J. Hunt, and M. Zhou, "Diffusion policies as an expressive policy class for offline reinforcement learning," in The Eleventh International Conference on Learning Representations.

[60] A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision making?" in The Eleventh International Conference on Learning Representations.

[61] H. Chen, C. Lu, C. Ying, H. Su, and J. Zhu, "Offline reinforcement learning via high-fidelity generative behavior modeling," in The Eleventh International Conference on Learning Representations.

[62] M. Reuss, Ö. E. Yağmurlu, F. Wenzel, and R. Lioutikov, "Multimodal diffusion transformer: Learning versatile behavior from multimodal goals," in First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024.

[63] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in International Conference on Machine Learning. PMLR, 2022, pp. 9902–9915.

[64] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 5923–5930.

[65] J. Carvalho, A. Le, M. Baierl, D. Koert, and J. Peters, "Motion planning diffusion: Learning and planning of robot motions with diffusion models," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023.

[66] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748–8763.

[67] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," arXiv preprint arXiv:2402.05672, 2024.

[68] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4195–4205.

[69] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in International Conference on Learning Representations, 2018.

[70] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 12, pp. 4217–4228, 2021.

[71] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," Advances in neural information processing systems, vol. 30, 2017.

[72] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," Advances in neural information processing systems, vol. 35, pp. 23 192–23 204, 2022.

[73] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," arXiv preprint arXiv:2402.10329, 2024.

[74] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," arXiv preprint arXiv:2410.18647, 2024.

[75] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in Forty-first International Conference on Machine Learning, 2024.

[76] M. Zhu, Y. Zhu, J. Li, J. Wen, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng, F. Feng, et al., "Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation," arXiv preprint arXiv:2409.14411, 2024.

[77] S. Dasari, O. Mees, S. Zhao, M. K. Srirama, and S. Levine, "The ingredients for robotic diffusion transformers," arXiv preprint arXiv:2410.10088, 2024.

[78] 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.

[79] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.

[80] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., "Lora: Low-rank adaptation of large language models," in International Conference on Learning Representations.

[81] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," arXiv preprint arXiv:2410.00425, 2024.