1. Abstract

In this appendix, we complement the main paper with additional analyses and results for AFRO. We first situate our approach within the broader literature on robotic manipulation. We then provide qualitative visualizations of AFRO rollouts on four real-world Franka tasks using point-cloud observations, as well as the complete per-task results on all Adroit and MetaWorld benchmarks. Next, we describe in more detail how we process the large-scale RH20T dataset to obtain point-cloud inputs for pre-training. Finally, we discuss current limitations of AFRO and outline several promising directions for scaling, enriching semantics, incorporating multi-view dynamics, and improving latent-space learning objectives.

2. Additional Related Work

2.1. Robotics Manipulation.

Robotic manipulation aims to transform object states in the physical world, from basic grasping[13, 14] to long-horizon, contact-rich [11] tasks. Early systems decomposed perception, grasp planning, and motion planning, while recent work learns end-to-end policies from rich sensory inputs such as RGB(-D) [1], tactile [5, 12, 15], and force feedback [10], often with additional safety constraints [9] or language conditioning. Vision–language–action models [6] further condition policies on natural-language goals but are still mostly built on 2D image encoders, which struggle with occlusions and precise 6-DoF reasoning. To overcome these limitations, recent approaches use 3D-aware policies that operate on depth, multi-view RGB-D, or point clouds, such as methods that infer volumetric features or explicit 3D action maps [3, 4, 7]. Our work is complementary: instead of introducing another policy head, AFRO provides a 3D dynamics-aware representation that can be plugged into existing diffusion-based manipulation policies with minimal changes.

3. Additional Real-World Visualizations

The main paper evaluates AFRO on four real-robot tasks that span non-prehensile pushing, precise contact interaction, and long-horizon pick-and-place motions using a Franka Emika arm and a top-down depth sensor.In Fruit Pick-and-Place, the robot must reach for a fruit at a randomized pose and place it into a distant basket; in Bell Pressing, it must localize a small bell and press it with sufficient accuracy to trigger the mechanism; in Block-to-Block Alignment, it pushes a movable block until its edge is aligned with a fixed reference block; and in Cover Block, it lifts a cup and places it so that the cup fully covers a target block. In all cases, object positions and orientations are randomized within bounded ranges to test robustness.

Figure 1 provides additional qualitative insight into how these tasks appear in the point-cloud observation space. For each task we render four representative frames, showing the pre-contact configuration, the approach of the end-effector, the main interaction phase (pressing, pushing, or grasping), and the final configuration after successful completion. The sequences illustrate the substantial 3D motion, self-occlusion, and depth noise present in real scenes, as well as the diversity of object geometries (e.g., thin bell handle versus bulky fruit or cup). AFRO is pre-trained and fine-tuned directly on such sequences, and the strong real-world performance reported in the main paper indicates that its latent dynamics modeling can extract task-relevant structure from these raw point-cloud trajectories.

4. Complete MetaWorld Results

Table 1 reports per-task success rates on all 16 simulated manipulation tasks used in our study: two Adroit hand tasks (Door, Pen) and fourteen MetaWorld tasks spanning four difficulty levels (Easy, Medium, Hard, Very Hard). For each method, we train a diffusion policy on top of the corresponding visual encoder using 100 expert trajectories for each Adroit task and 25 trajectories for each MetaWorld task, and evaluate success over 50 rollouts every 10 epochs; the table shows the best success rate over multiple runs under this shared protocol. Compared to all baselines, AFRO attains the highest mean performance and achieves the best result on the majority of tasks, ranking second only on Bin Picking, Coffee Pull, and Soccer. These fine-grained results complement the aggregated comparisons in the main paper and confirm that AFRO consistently improves over both 2D and 3D pre-training baselines across a broad range of contact-rich skills.

5. Further Details on RH20T

RH20T[2] (Robot–Human demonstration in 20TB) is a large-scale, real-world robotic manipulation dataset designed to support one-shot imitation learning and general skill learning across diverse, contact-rich tasks. The dataset contains over 110,000 robot manipulation sequences paired with an equal number of human demonstration videos, resulting in more than 50 million image frames in total and about 20 TB of data. It demonstrates over 140 different tasks, including a wide variety of operational tasks. Each robot sequence is recorded in the real world and includes synchronized multi-modal information such as RGB and depth images, force/torque measurements, audio, and low-level robot state and action signals, together with a corresponding human demonstration video and a language description for the same skill. We generate a point cloud for each frame by back-projecting the depth map into 3D space using the corresponding camera intrinsics. We then apply farthest point sampling (FPS) to downsample the raw point cloud to 1024 points. Empirically, we find that 1024 points provide sufficient geometric coverage to capture the object contours while at the same time substantially reducing the computational cost of the PointTransformer encoder.

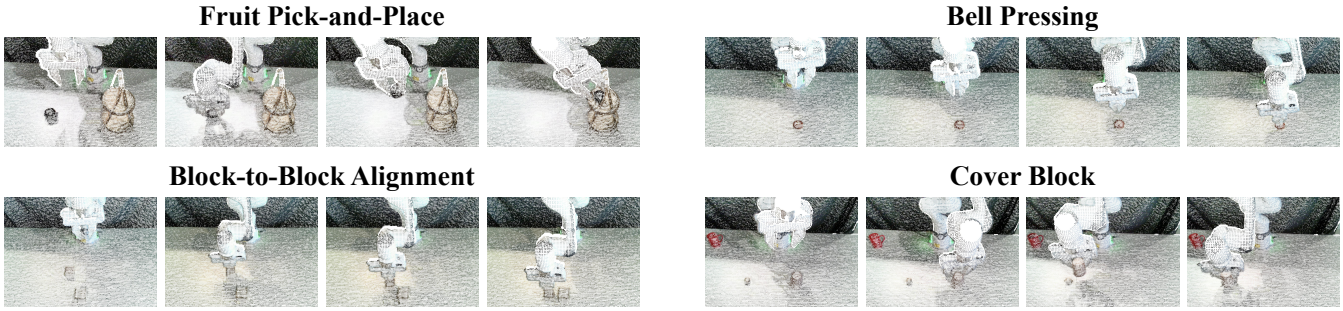

Figure 1. **Point-cloud rollouts for four real-world tasks.** We visualize the 3D point clouds captured by the top-down RealSense L515 depth camera for four representative manipulation tasks on the Franka platform: Fruit Pick-and-Place, Bell Pressing, Block-to-Block Alignment, and Cover Block (from left to right and top to bottom). Each block shows four temporally ordered point clouds from a successful AFRO rollout, illustrating the evolution from the initial configuration, through approach and contact, to task completion. Across tasks, the manipulated objects (fruit and basket, bell, blocks, cup and target block) undergo large spatial motion while the surrounding table and robot geometry remain largely static. These visualizations highlight that AFRO is trained directly on such raw point clouds and must learn dynamics-aware representations that are sensitive to object motion and interaction, yet robust to background clutter and viewpoint changes.

Table 1. **Comparison across 16 simulated manipulation tasks.** Success rates (%) for AFRO and competing methods on two Adroit tasks and fourteen MetaWorld tasks grouped by difficulty. Best results are highlighted in red, and second-best results are highlighted in orange. All methods use the same diffusion-policy architecture and training protocol: policies are trained from 100 expert trajectories on Adroit and 25 trajectories on each MetaWorld task, and evaluated over 50 rollouts every 10 epochs; we report the best success rate over multiple runs.

	Adroit		**MetaWorld (Easy)**			**MetaWorld (Medium)**		
Method	Door	Pen	Dial Turn	Handle Press	Peg Unplug Side	Bin Picking	Coffee Pull	Peg Insert Side
CILP	61	84	0	88	0	0	72	0
DINOv2	76	84	6	90	6	0	58	0
PointNet	80	72	56	90	68	62	66	54
PointMAE	64	76	60	88	84	16	94	82
PointDif	76	78	76	90	84	18	90	84
Dynamo	76	68	18	84	24	18	34	14
Dynamo-3D	73	76	70	98	86	18	68	76
FVP	66	76	72	88	80	16	66	70
DP3	70	80	76	86	86	18	90	74
AFRO (Ours)	82	84	78	98	88	20	92	92

	MetaWorld (Medium)			**MetaWorld (Hard)**		**MetaWorld (Very Hard)**		
Method	Push Wall	Soccer	Sweep	Pick Out of Hole	Push	Stick Pull	Stick Push	Pick Place Wall
CILP	6	0	6	0	10	66	100	0
DINOv2	2	24	0	0	10	66	100	0
PointNet	40	24	52	10	36	30	88	48
PointMAE	66	22	88	16	68	58	76	76
PointDif	38	26	92	20	62	52	76	48
Dynamo	26	26	12	16	12	16	72	14
Dynamo-3D	58	24	88	4	78	60	100	80
FVP	36	34	44	26	42	24	84	78
DP3	78	38	92	24	74	58	88	94
AFRO (Ours)	80	36	98	32	78	78	100	94

6. Limitations and Future Work

Although AFRO achieves strong performance across diverse simulated and real-world manipulation tasks, several limitations remain.:contentReference[oaicite:0]index=0 First, our current pre-training regime is moderate in scale compared with truly large vision foundation models: we use task-specific simulation data and a subset of RH20T rather than hundreds of millions of frames. Second, the objective is predominantly dynamics-driven. While this helps encode transition structure, the learned features still lack the rich se-

mantic coverage that large 2D or 3D visual models obtain from web-scale data. Third, our framework is instantiated with single-view point clouds from a fixed depth camera; it does not yet exploit the multi-view nature of real 3D environments, such as fusing observations from multiple cameras over time. Fourth, the latent action is an abstract variable optimized only for predictive performance; it does not have explicit physical meaning (e.g., end-effector displacement, velocity, or contact force), which limits its interpretability and potential reuse. Finally, AFRO relies on VICReg-style variance regularization to prevent collapse in latent space; as shown in our ablations, removing this constraint severely harms performance, indicating that training stability is still tied to a relatively strong handcrafted objective.

These observations suggest several directions for future work. (1) **Scaling and semantic grounding.** We plan to combine AFRO with strong semantic backbones such as VGGT [8] or DINOv2 style models, either by distilling their features into the 3D encoder or by joint multi-task pre-training. This would aim to obtain representations that are simultaneously dynamics-aware and semantically rich. (2) **Larger-scale pre-training.** A natural extension is to train AFRO on substantially larger and more diverse 3D robot datasets, including multi-embodiment, multi-task, and web-scale synthetic data, to approach the scale of general-purpose visual foundation models for robotics. (3) **Multi-view 3D dynamics.** Incorporating multi-view inputs and explicitly modeling cross-view consistency of dynamics (e.g., by fusing trajectories from several cameras) could make the representation more robust to occlusion and viewpoint changes and better suited to mobile manipulation. (4) **Physically grounded latent actions.** We are interested in tying latent actions to physically interpretable quantities, for example by supervising them with estimated optical flow, 3D scene flow, or end-effector trajectories, so that the latent space captures meaningful motion primitives that can be reused across tasks and robots. (5) **Collapse-free objectives beyond VICReg.** Finally, we aim to explore alternative self-supervised objectives—such as JEPA-style prediction, information-theoretic regularization, or architectural constraints—that maintain feature diversity without relying on explicit variance penalties, potentially simplifying training and further improving robustness of the learned 3D representations.

References

[1] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44 (10-11):1684–1704, 2025. 1

[2] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023. 1

[3] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. *arXiv preprint arXiv:2306.17817*, 2023. 1

[4] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning (CoRL)*, pages 694–710. PMLR, 2023. 1

[5] Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025. 1

[6] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1

[7] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 1

[8] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 3

[9] Taowen Wang, Cheng Han, James Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6948–6958, 2025. 1

[10] Zijian Wang and Mac Schwager. Kinematic multi-robot manipulation with no communication using force feedback. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 427–432. IEEE, 2016. 1

[11] Qiwei Wu, Xuanbin Peng, Jiayu Zhou, Zhuoran Sun, Xiaogang Xiong, and Yunjiang Lou. Rttf: Rapid tactile transfer framework for contact-rich manipulation tasks. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2913–2920. IEEE, 2024. 1

[12] Qiwei Wu, Haidong Wang, Jiayu Zhou, Xiaogang Xiong, and Yunjiang Lou. Tars: Tactile affordance in robot synesthesia for dexterous manipulation. *IEEE Robotics and Automation Letters*, 2024. 1

[13] Lixin Xu, Zixuan Liu, Zhewei Gui, Jingxiang Guo, Zeyu Jiang, Zhixuan Xu, Chongkai Gao, and Lin Shao. Dexsingrasp: Learning a unified policy for dexterous object singulation and grasping in cluttered environments. *arXiv preprint arXiv:2504.04516*, 2025. 1

[14] Yuhang Zheng, Xiangyu Chen, Yupeng Zheng, Songen Gu, Runyi Yang, Bu Jin, Pengfei Li, Chengliang Zhong, Zengmao Wang, Lina Liu, et al. Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping. *IEEE Robotics and Automation Letters*, 2024. 1

[15] Jiayu Zhou, Qiwei Wu, Jian Li, Zhe Chen, Xiaogang Xiong, and Renjing Xu. Gentle manipulation policy learning via demonstrations from vlm planned atomic skills. *arXiv preprint arXiv:2511.05855*, 2025. 1