

Simplifying 3D Robot Policy Setup with 3D Foundation Models

Cameron Smith

University of Southern California
cosmith@usc.edu

Toan Nguyen

University of Southern California
tientoan@usc.edu

Abstract

While 3D representations offer significant advantages for robot policy learning in terms of data efficiency, robustness, and generalization, existing approaches face a critical practical barrier: the reliance on expensive, noisy depth sensors such as LiDAR or stereo cameras. In this work, we explore whether recent 3D foundation models, which can reconstruct high-quality 3D point clouds directly from unposed RGB images, can serve as a practical replacement for traditional depth sensing hardware. To investigate this question, we first conduct systematic experiments in simulation to figure out the optimal 3D policy learning architecture. Our results on the LIBERO-Object benchmark reveal that lifting 2D foundation models’ features to 3D significantly outperforms raw RGB representations, and that the lightweight DP3 encoder consistently surpasses the more complex PointNet architecture. Building on these insights, we demonstrate preliminary real-world results showing that foundation-model reconstructions from single RGB images can enable effective policy learning for robotic grasping tasks. Our findings suggest a promising path toward more accessible 3D robot learning systems that eliminate the need for specialized depth sensing hardware while maintaining strong performance.

1 Introduction

3D robot learning is quickly gaining traction within the robot learning community, driven by its potential to provide richer spatial understanding than 2D approaches, offering notable advantages in data efficiency, robustness, and generalization [18, 16, 1]. However, developing and deploying 3D policy learning approaches in practice still encounter significant sources of friction. A major source of friction lies in the need to obtain accurate 3D scene information, which typically requires specialized hardware and careful calibration. In particular, existing 3D robot policy learning approaches rely heavily on depth sensors such as LiDAR or stereo cameras, which are often expensive, noisy, and cumbersome to integrate into everyday robotic setups. Meanwhile, a new generation of 3D foundation models [12, 15, 13, 14] has emerged, demonstrating the ability to infer high-quality 3D point clouds directly from unposed RGB images, offering a compelling alternative to traditional pipelines that rely on depth cameras. In this work, we investigate whether these foundation-model reconstructions can serve as a practical replacement for depth sensors in 3D policy learning. Our goal is to determine if policies trained on foundation-model geometry can match or even surpass those trained on depth-sensor-based 3D representations. Preliminary results show that with 3D foundation models, we can get reasonably high-quality 3D point clouds from a single RGB image, and that the model using these point clouds shows promising results in a grasping task using a low-cost SO-100 robot [11].

Another challenge in 3D policy learning is the absence of consensus on the most effective 3D representation or the optimal point cloud encoder to use when designing 3D policy learning methods. To address this question and also to come up with a strong 3D policy architecture, we conduct systematic experiments in simulation to evaluate different combinations of two point cloud representations (xyz coordinates with raw RGB colors, and xyz coordinates with lifted features from 2D CLIP [8] foundation models) and two point cloud encoders (PointNet [7] and a lightweight yet carefully designed DP3 encoder [16]). Experimental results on the LIBERO-Object benchmark [5] show that the CLIP features are significantly better than raw RGB inputs, and that the DP3 encoder largely outperforms PointNet despite its simplicity. These results directly inform our model development in the real-world setting.

Our findings can be summarized as follows:

- 2D foundation-model features can greatly benefit 3D robot policy learning.
- Despite its relatively simple design, the DP3 point cloud encoder consistently outperforms the larger and more sophisticated PointNet model.
- Leveraging 3D foundation models to reconstruct high-quality point clouds from single 2D images is feasible and yields encouraging preliminary results in real-world robotic settings.

2 Related Work

3D Policy Learning. Recent advances in robot manipulation have increasingly leveraged 3D visual representations for learning visuomotor policies from demonstrations. Recent work has shown that 3D visual representations, particularly point clouds, offer significant advantages over 2D RGB representations in terms of generalization and robustness to viewpoint changes [6, 18]. Voxel-based approaches like PerAct [9] encode RGB-D observations into voxelized scenes and use Perceiver Transformers to predict discretized actions, achieving remarkable data efficiency across multi-task scenarios. Building on this, 3D Diffuser Actor [1] lifts 2D features to 3D space and employs diffusion models for action generation, demonstrating state-of-the-art performance on manipulation benchmarks. Similarly, 3D Diffusion Policy [16] combines 3D visual representations with a diffusion policy model to achieve strong performance across diverse manipulation tasks. More recently, flow-matching-based methods like FlowPolicy [17] have emerged as alternatives to diffusion models, offering more stable training while maintaining high performance with point cloud observations. Despite these significant progresses, existing 3D policy learning methods predominantly rely on depth cameras to capture point cloud or voxel representations, which are expensive, noisy, and add hardware complexity to real-world robotic system deployment. In this work, we explore an alternative approach that leverages 3D foundation models to reconstruct point cloud scenes from 2D RGB images captured by standard cheap cameras, eliminating the need for dedicated depth sensors.

3D Scene Reconstruction. The emergence of 3D foundation models has enabled scene reconstruction from monocular or sparse multi-view RGB images. DUS3R [15] pioneered a paradigm shift by directly predicting 3D point maps from image pairs without requiring known camera parameters, eliminating the complex multi-stage pipelines of traditional Structure-from-Motion and Multi-View Stereo approaches. Building on this foundation, MAST3R [4] extended DUS3R with dense local feature matching capabilities and metric 3D reconstruction, demonstrating strong accuracy in establishing pixel correspondences across thousands of images. More recently, VGGT [12] achieved state-of-the-art performance by predicting all key 3D scene attributes, including camera parameters, depth maps, point maps, and 3D tracks—from one to hundreds of views in a single feed-forward pass under one second, significantly outperforming optimization-based methods. For monocular geometry estimation, MoGe [13] introduced affine-invariant point map representations with novel supervision strategies, achieving remarkable generalization across diverse datasets, while MoGe-2 [14] further extended this to metric-scale 3D reconstruction with fine-grained detail recovery. Despite these impressive advances in reconstructing 3D scenes from RGB images, applying these 3D foundation models to enable robot policy learning remains a largely unexplored research direction, presenting significant opportunities for reducing hardware costs and complexity in robotic systems.

3 Optimal 3D Policy Architecture

Despite significant advances having been made in 3D robot learning, there is still no unified consensus on what the optimal architecture design is for 3D policy learning models. To make an informed choice when designing the 3D policy architecture for our real robot, we first aim to answer two key questions via experiments in simulation: (1) What is the most effective point cloud representation for policy learning? and (2) What is the most effective point cloud encoder for 3D policy learning?

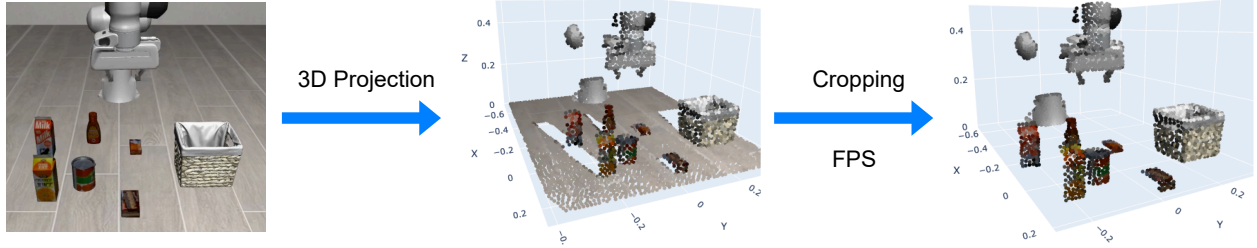


Figure 1: Summary of the point cloud extraction process.

Point Cloud Representation. We experiment with two point cloud representations. The first uses raw 3D coordinates with RGB colors. The second representation combines 3D coordinates with lifted features from 2D foundation models (e.g., CLIP [8], DINOv3 [10]) derived from the corresponding 2D images [8]. To construct this representation, we retain the RGB images associated with each point cloud, along with the pixel location corresponding to each point. We then apply the CLIP ViT encoder to extract a patch-based feature map from the image and assign to each point the feature vector of the patch containing its corresponding pixel. The main purpose of experimenting with this representation is to validate the usefulness of features from 2D foundation models in 3D robot learning.

Point Cloud Encoder. We also experiment with different point cloud encoders. Specifically, we benchmark the two point cloud encoders, which are the PointNet encoder [7] and the DP3 encoder introduced in 3D Diffusion Policy [16]. The PointNet encoder has roughly 2.7M parameters, while the DP3 point cloud encoder is lightweight and conceptually simple with only about 59K parameters. However, in experiments in [16], it obtained significantly higher performance than other widely-used point cloud networks. In our work, we aim to verify the effectiveness of this point cloud encoder.

Data Preparation and Experiment Setting. We use the LIBERO-Object task suite of the LIBERO benchmark [5] for our experiments. The task suite consists of 10 language-driven pick-and-place tasks performed by a Franka Panda robot. Each task comes with 50 high-quality human demonstrations. We first follow OpenVLA [3] to remove no-op actions that do not change the robot’s state and remove episodes that are unsuccessful. We then extract the point cloud observations from processed episodes. In particular, from a 2D image, we use the provided depth map and camera parameters to extract a point cloud of the whole scene in the world frame. We then crop the point cloud to retain only the points within the robot’s workspace, and remove floor points to produce a point cloud that contains only the robot and relevant objects. Finally, we employ the farthest point sampling (FPS) algorithm to obtain a point cloud of 2048 points. The point cloud extraction process is summarized in the Figure 1.

We design three model variants with different combinations of point cloud representations and point cloud encoders: (i) 3D coordinates with RGB colors and PointNet, (ii) 3D coordinates with RGB colors and DP3, and (iii) 3D coordinates with CLIP features and DP3. To facilitate fair comparisons, we keep the policy backbone the same across model variants. In particular, we leverage the transformer policy model introduced in the original LIBERO paper [5], but replace the visual encoder with the point cloud encoders, all of which output a 64-D compact embedding of the whole point cloud scene. We train all models in 50 epochs and set the batch size to 32. For evaluation, we perform 20 rollouts with different initial object positions for each of the 10 tasks (200 rollouts in total for each model variant) and report the overall success rate across all rollouts. The maximum number of steps in rollout is 200. After 200 steps, if the task is still not completed, the episode is marked as a failure.

Results. Table 1 clearly shows that the combination of CLIP features with the DP3 encoder dramatically outperforms the other configurations, reaching over 90% success. This confirms the effectiveness of leveraging 2D foundation models’ features in 3D policy learning. The second-best variant is the model that uses CLIP features and the DP3 encoder, but it is still far behind, while the model that uses PointNet completely

Table 1: Simulation experimental results.

Representation	Encoder	↑ Success Rate
3D Coordinates with RGB	PointNet	0.0
3D Coordinates with RGB	DP3	<u>18.8</u>
3D Coordinates with CLIP	DP3	91.3

Task: "Pick up the alphabet soup and place it in the basket"

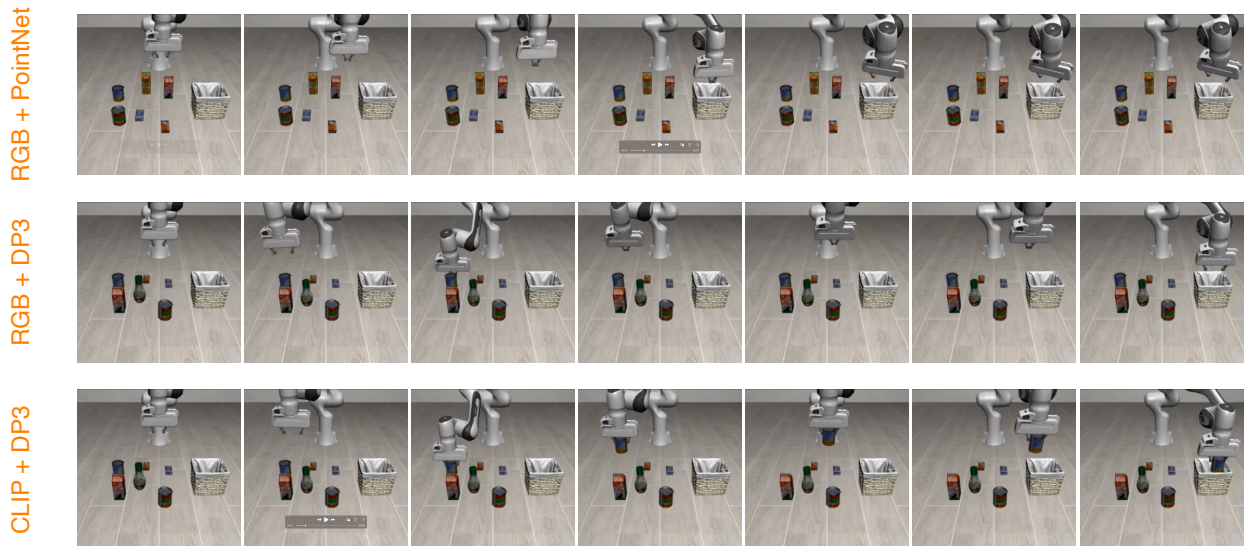


Figure 2: Rollout examples.

fails in every episode. To further analyze the performance of the three model variants, we also examine their actual rollouts, shown in Figure 2. From these rollouts, we observe that the PointNet-based model performs poorly across all tasks. The model using RGB colors and DP3 exhibits reasonable motion toward the target object but consistently fails to grasp it, resulting in a low success rate. In contrast, the model using CLIP features and DP3 successfully grasps the object and subsequently places it into the basket. Based on the simulation results, we adopt the 2D foundation-model-enhanced point cloud representation and the DP3 encoder for our real-world experiments, but replace CLIP with DINOv3 [10] because it is generally a more updated and capable model.

4 Using 3D Foundation Models for 3D Robot Scene Representations

Inferring the 3D Point Cloud. Traditional 3D robot policy pipelines use depth sensors to infer the 3D point cloud as a scene representation; however, these sensors are expensive, empirically exhibit firmware issues, and produce sparse and noisy depth estimates. Instead, we propose to leverage 3D foundation models for the 3D point cloud. We experimented with using MapAnything [2], a 3D prediction method conditioned on camera poses – with the robot poses as input, but found the pose conditioning to be noisy and finicky on the scale of the scene. We then tried unposed VGGT [12] with two input views and found the reconstructions to be compelling. To make the setup even simpler, we tried the single-view method MoGe [13] and found the reconstructions also sufficient, so we proceeded with that to simplify the setup from needing multiple cameras to just one.

Transforming the Point Cloud to the Robot Coordinate Frame. The main challenge in leveraging point clouds from 3D foundation models is that their metric scale and coordinate system are arbitrary,

RGB frame to 3D Robot State

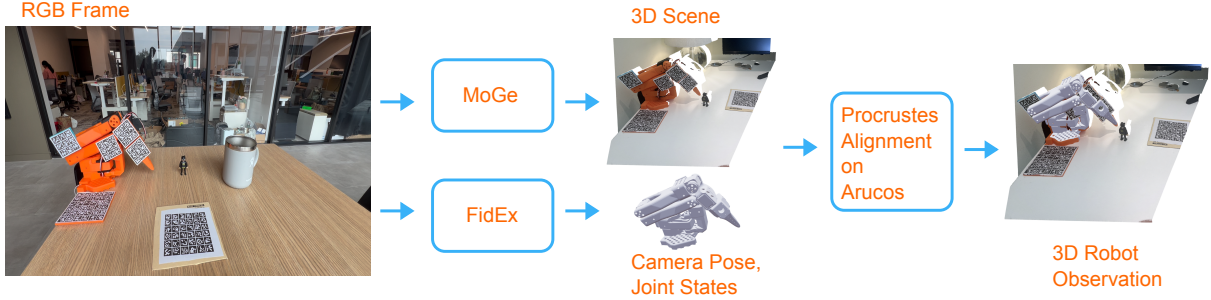


Figure 3: An overview of how we map a single RGB demonstration to a 3D Robot demonstration.

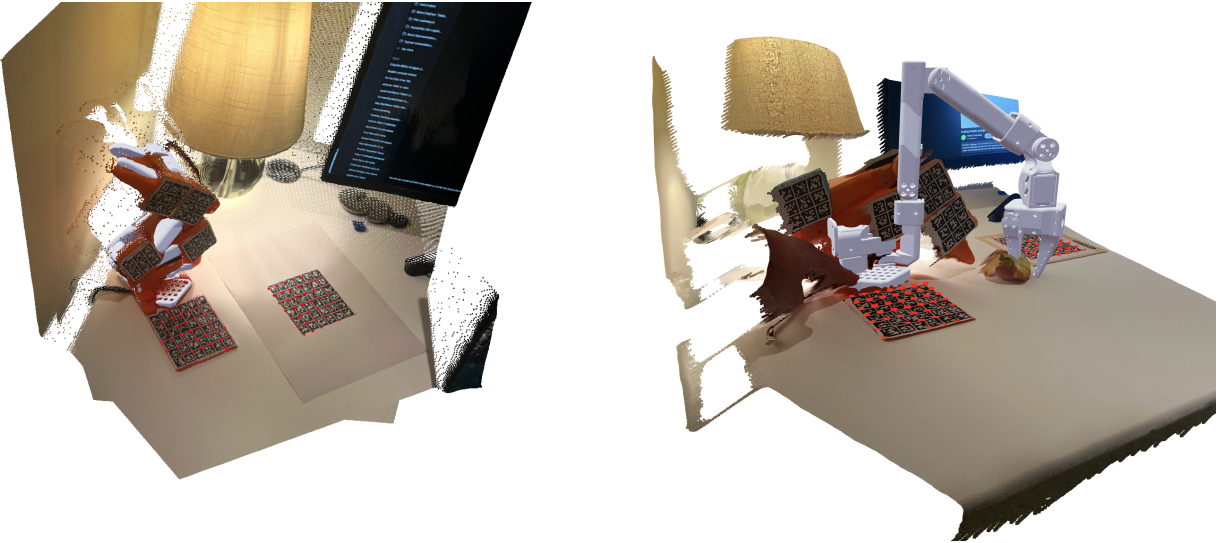


Figure 4: Supporting ArUco Board for Alignment: In addition to the ArUco board for the robot base inference, we add a second ArUco board in the scene. We use the metric points from both boards for Procrustes alignment from the inferred point cloud and the robot frame.

Sample Dataset Robot-Object Alignments

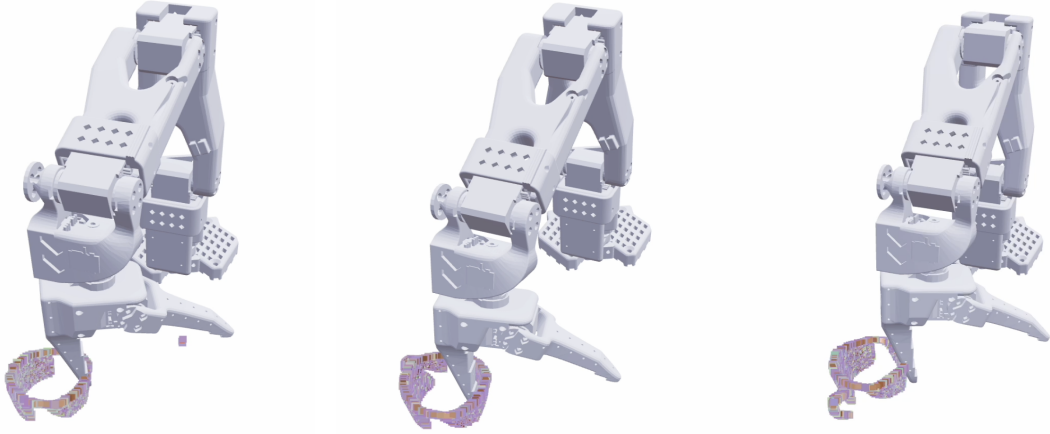


Figure 5: **Qualitative mug alignment examples.** Three example scenes showing the reconstructed mug and robot in the aligned 3D coordinate frame. In the left and center examples, the mug and gripper geometry are well aligned with the point cloud and ArUco boards. In the right example, the gripper slightly penetrates the mug, corresponding to a residual error of roughly 0.5cm between the inferred cup position and the true object boundary, illustrating a common failure mode of our current alignment pipeline with a margin of error in the estimated geometry.

Table 2: Real-world experimental results.

	Train	Test
End-Effector Position Error (mm)	13.74	30.74
End-Effector Orientation Error (degree)	4.83	4.37

whereas we require the point cloud in the robot frame and of metric scale (need to recover the SIM3 transformation). We use FidEx (an in-submission project from our group) to obtain the camera pose from a single frame, which we can use to transform the points to the correct coordinate system, but the scale is still arbitrary. To resolve this, we do a simple procrustes alignment on the metric ArUco points from the FidEx ArUco board and the 3D points predicted from the foundation model. We found that this alignment is good but can be greatly enhanced by adding a second ArUco board in the scene, and using these metric points for alignment as well.

Real-World Dataset and Policy Training. After validating the 3D alignment quality, we collected a real-world dataset on a mug grasping task with the SO-100 robot [11]. For each demonstration, we record the robot joint states and end-effector pose at the grasp time, together with the 3D point cloud reconstructed by MoGe [13]. We then crop the point cloud to a fixed workspace region and lift DINOv3 features [10] into 3D by using the corresponding transformed pointmap coordinate.

Qualitatively, the resulting geometry and features are largely consistent with the rendered robot and object states (Figure 5 and Figure 6). The mug, table, and robot base generally align well with the ArUco boards and the inferred 3D geometry, and most grasps appear visually plausible. However, there are still noticeable imperfections at contact boundaries: in some examples, the gripper slightly penetrates the mug (Figure 5, right), corresponding to a residual pose error of roughly 0.5cm between the inferred cup position and the true object boundary. These small but important misalignments highlight the limit of our current monocular reconstruction and alignment pipeline.

Sample Dataset Validation Predictions

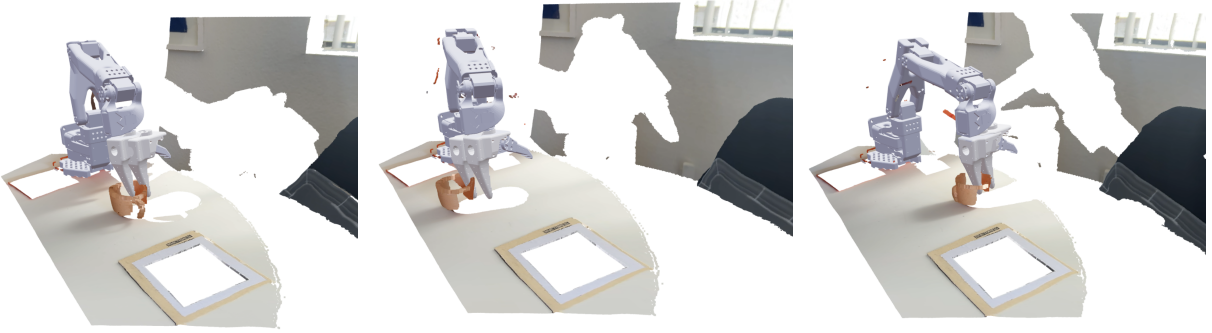


Figure 6: **Real-world grasp prediction results.** Qualitative results of the learned DP3-style policy with DINOv3-augmented point clouds. For each example, we show the ground-truth grasp pose and the predicted end-effector pose. The model predictions are usually close—often within a few centimeters and a few degrees of the target—but small residual errors remain, consistent with our quantitative results ($\approx 13.7\text{mm}$ train vs. 30.7mm test position error). These errors are large enough to cause missed or unstable grasps, suggesting that further improvements in data efficiency and policy formulation are needed for reliable deployment.

Using this dataset, we train a DP3-style point-cloud policy, informed by our simulation architecture study in Section 3. The policy takes the cropped DINO-augmented point cloud as input and predicts the 6-DoF grasp pose of the end effector. We collect a small dataset of 30 demonstrations in total, using 25 for training and 5 for held-out evaluation.

Results. Prediction error results of the trained policy on training and testing data are shown in Table 2. Figure 7 visualizes representative predictions from this model. While quantitative results indicate that the network learns a reasonable mapping from foundation-model point clouds to grasp poses, a test error of $\approx 3\text{cm}$ in position is not reliable enough for deployment on our mug-grasping task, where a few millimeters of error can lead to missed grasps.

5 Discussion

While our method presents commendable preliminary results, it still needs a lot of future improvements. We hypothesize that the main bottlenecks of the current setting are (i) the small amount of real-world data collected and (ii) the difficulty of training point-cloud-based policies, which are known to be data-hungry and often exhibit poor generalization compared to image-based counterparts, and (iii) the precision of alignment in the monocular-estimated vs real-world geometry. In ongoing work, we are exploring a more data-efficient gripper prediction formulation that uses local image and coordinate predictions rather than global point-cloud-action representations. After getting sufficiently good results on the mug-grasping task, we plan to further improve our policy architecture and benchmark it on diverse and more challenging tasks.

6 Conclusion

In this work, we investigated whether 3D foundation models can simplify the deployment of 3D robot policies by replacing traditional depth sensors with point clouds inferred directly from RGB images. Through controlled simulation experiments, we identified an effective 3D policy architecture that combines 2D foundation-model features with the lightweight DP3 encoder, yielding strong performance and guiding our real-world design choices. Our preliminary real-world experiments demonstrate that monocular 3D reconstructions, when aligned to the robot frame and enriched with DINOv3 features, can support meaningful

Sample Dataset Robot-Scene Alignments

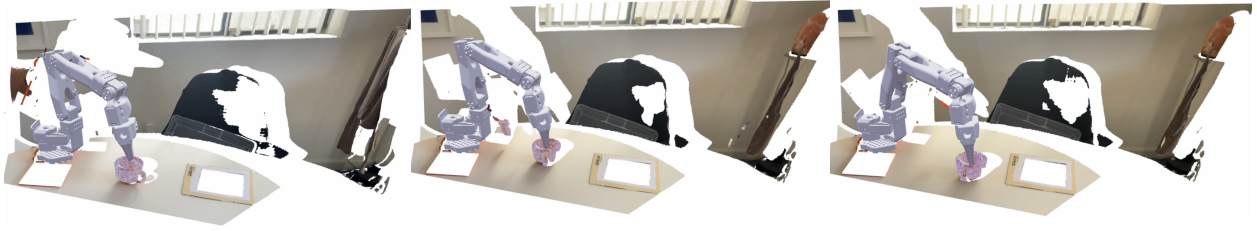


Figure 7: **Overall robot-scene alignment quality.** Additional examples of real-world scenes after alignment, showing the robot, table, and mugs overlaid on the 3D foundation-model point cloud. Across multiple viewpoints and object configurations, the reconstructed geometry is largely consistent with the physical setup, and the robot links line up well with the inferred table and ArUco markers, indicating that our calibration and alignment procedure is generally robust.

grasp-pose prediction, though residual geometric errors still limit task-level success. Overall, our results highlight a promising path toward accessible, low-cost 3D robot learning pipelines powered by foundation-model geometry, while also revealing key challenges in alignment accuracy, data efficiency, and policy robustness that future work must address.

References

- [1] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *Conference on Robot Learning*, pages 1949–1974. PMLR, 2025.
- [2] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kongschieder. MapAnything: Universal feed-forward metric 3D reconstruction, 2025. arXiv preprint arXiv:2509.13414.
- [3] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.
- [4] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.
- [5] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.
- [6] Skand Peri, Iain Lee, Chanh Kim, Li Fuxin, Tucker Hermans, and Stefan Lee. Point cloud models improve visual robustness in robotic learners. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17529–17536. IEEE, 2024.
- [7] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas 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*, pages 652–660, 2017.

- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [9] 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.
- [10] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.
- [11] TheRobotStudio. So-arm100 — standard open robot arm 100. <https://github.com/TheRobotStudio/SO-ARM100>, 2024. Open-source 6-DOF robotic arm kit.
- [12] 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.
- [13] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025.
- [14] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025.
- [15] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.
- [16] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.
- [17] Qinglun Zhang, Zhen Liu, Haoqiang Fan, Guanghui Liu, Bing Zeng, and Shuaicheng Liu. Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14754–14762, 2025.
- [18] Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *Advances in Neural Information Processing Systems*, 37:77799–77830, 2024.