A_0: An Affordance-Aware Hierarchical Model for General Robotic Manipulation

Supplementary Material

A. Social Impact

The internet datasets we used, PixMo-Points, HOI4D, and the real-robot dataset Droid, are all publicly available and transparent. The ManiSkill-5k dataset was collected in a simulation environment, while the real-robot datasets we collected do not contain any personal information. We plan to release these datasets in the future. Our method has no ethical risk on dataset usage and privacy violation since all the benchmarks are publicly available and transparent.

B. Limitations and Future work

B.1. Limitaions

Our method has two main limitations:

- Our model relies on methods like gripper samplers to predict grasp poses for action execution. However, existing approaches in this area often exhibit suboptimal performance and limited generalization across different tasks.
- Our method requires a depth map to estimate height, followed by refinement using a VLM. However, this approach may not perform well in tasks involving occluded objects.

B.2. Discussion

- **Orientation-sensitive tasks.** Our method focuses on high-level spatial understanding. The action execution module is fully modular and can be replaced. For tasks requiring precise orientation and nuanced 6D manipulation—such as liquid pouring and revolute-drawer opening—We will choose an optimal observation viewpoint that allows A_0 to predict the necessary waypoints. As demonstrated in prior work (e.g., ATM: Any-point trajectory modeling [6]), a track-guided policy can be trained with only a few demonstrations to achieve accurate control. For more complex tasks, VLMs like GPT-4o can be employed to decompose the task into a sequence of subtasks and A_0 can address each step individually.
- **Long-horizon planning.** Our method faces this common limitation shared by affordance-based and modular approaches. Even current VLA models struggle with long-horizon tasks. In future work, we will address this by leveraging VLMs such as GPT-4o to decompose long-horizon tasks into a sequence of shorter subtasks (e.g., $\pi_{0.5}$), which can then be executed stage-by-stage using A_0. While MOKA and ReKep handle multiple objects via prompt VLM, we have incorporated VLM for high-level planning in our new experiments. This enables our model to perform more complex tasks, such as inserting flowers into the bottle and putting fruits on the plate, as shown in Fig. 2.

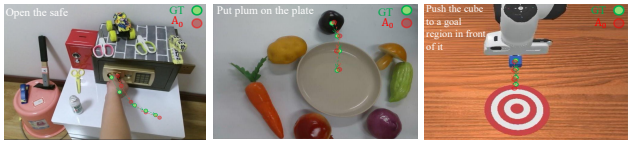


Figure 1. A_0 can predict accurately on different datasets. These three images are sourced from Droid, our Franka robot, and the ManiSkill simulation environment.

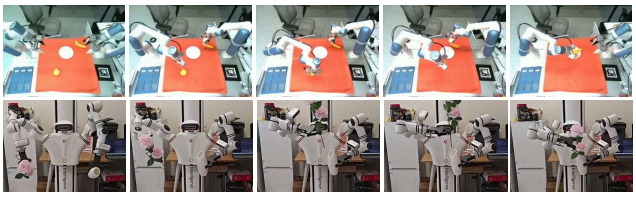


Figure 2. We incorporate a VLM for high-level planning, which enables our model to perform more complex tasks. First row: put fruits on a plate. Second row: insert flowers into the bottle.

B.3. Future Work

Based on our analysis of the limitations of our method, we plan to pursue the following future work:

- First, improving grasp pose estimation: Our current method relies on a gripper sampler to obtain grasp poses. A potential improvement could be leveraging a VLM to visually assist in selecting the best gripper position from gripper candidates or directly prompting the VLM to generate a grasp pose.
- Second, improving height estimation: Currently, the grasp height is obtained by prompting an LLM. We can refine our model by incorporating depth, gripper length, and other relevant information as conditions to directly predict the height.

C. Dataset

We have collected data from four types of sources: internet data, real-robot data, human-centric data, and simulator data. Below is a detailed description of each of these four types of data.

Contact-point Localization Data: We select one million samples with only a single coordinate point (contact point c_0^{2D}) annotation from the PixMo-Points [1] dataset,

named PixMo-One-Point. Every sample consists of one image, object label and corresponding coordinate.

Real Robotic Data: We developed a multi-stage annotation pipeline to capture high-quality manipulation trajectories from real robotic interactions. Human annotators identified target objects and initial contact points in videos, followed by automatic tracking using the CoTracker model [2] to generate trajectory waypoints. In the second method, we employed Molmo [1] to annotate the initial point and subsequently used SAM2 [5] to segment and track the object mask across frames. Each trajectory was manually verified for accuracy. This semi-automated approach produced a dataset of verified manipulation waypoints. We randomly selected and annotated 3,056 trajectory samples from the DROID [3] dataset, naming it DROID-3k.

Human-centric Data: Compared to robot interaction data, video-based human-object interaction data like HOI4D [4] are more accessible and semantically rich. The HOI4D dataset [4] encompasses 16 distinct object categories (e.g., toy cars, bottles) and covers 6 different tasks, including pick-and-place, opening a drawer, and pulling a toy car. Each category includes multiple videos, amounting to a total of 3,572 videos. Each video is comprised of several segmented actions, and by employing various action labels to delineate these segments, a total of 22,140 unique video segments were obtained. We convert the original dataset into a 2D waypoint format, where the center point of the object's 2D mask in each frame is used as the waypoint, and the combination of the action and object name serves as the instruction. We refer to the transformed dataset as HOI4D-22k.

Simulator Data: To adapt our model for various deployment environments, we collected 4965 trajectories from the ManiSkill Scene dataset, converting 3D data to 2D for compatibility. There are five tasks: "Peg Insertion Side", "Plug Charger", "Pull Cube Tool", "Push Cube", and "Stack Cube". Each task contains about a thousand trajectories. We named this dataset Maniskill-5k. Camera angles were adjusted to diversify scene data.

The image resolutions of the HOI4D, Maniskill, and DROID datasets are 1920×1080, 512×512, and 320×180, respectively.

D. Compare with Robopoint

To compare our approach with Robopoint, we evaluated both methods on two standard datasets: HOI4D and DROID. We measured the Mean Absolute Error (MAE) of the first predicted interaction pixel compared to the ground truth. For our model, we used the first waypoint directly, while for Robopoint, which predicts interaction regions rather than specific points, we used the average position of all predicted points for comparison.

As shown in Table 3, our method achieves substantially lower MAE scores across both datasets. Specifically, on HOI4D, our model achieves an MAE of 54.46 compared to Robopoint's 121.09, representing a 55.2% reduction in error. Similarly, on DROID, our approach attains an MAE of 14.13 versus Robopoint's 27.47, a 40.4% improvement. These results demonstrate that our method provides more precise interaction point predictions, which is crucial for accurate robot manipulation tasks.

Method	HOI4D (MAE)	DROID (MAE)
Robopoint	121.09	27.47
Ours	**54.46**	**14.13**

Table 1. Comparison of Mean Absolute Error (MAE) for the first interaction pixel between our method and Robopoint on HOI4D and DROID datasets. Lower values indicate better performance.

E. Real World Experiment

We conduct our real world robot experiments on multiple robot platforms, including Kinova, Franka, and Realman Robot, as shown in Figure 3, Figure 4 and Figure 5.

F. Annotation platform

Our annotation platform is a multi-source, semi-automated system designed to generate high-quality spatial affordance data for robotic manipulation. It integrates real robotic interactions, human-object interaction datasets, simulation environments, and large-scale internet-sourced datasets to create a standardized Embodiment-Agnostic Affordance Representation.

G. Additional Qualitative Result on Franka

As shown in Fig. 6, our approach generalizes well to different backgrounds on Franka Robot. The text below each raw of images serves as the instruction.

References

[1] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 2

[2] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv:2307.07635*, 2023. 2

[3] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen,

Figure 3. Real World Kinova Gen3 Robot

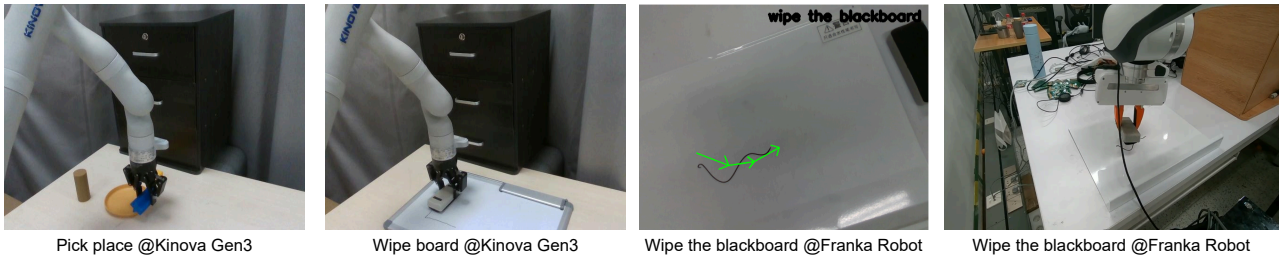


Figure 4. Real World Franka Robot

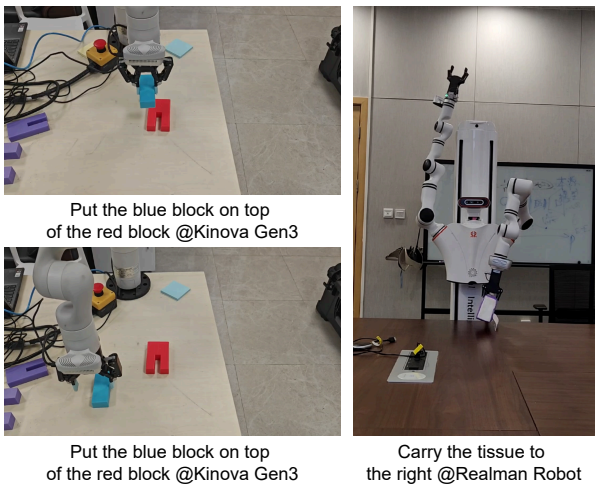


Figure 5. Real World Kinova Gen3 Robot and Realman Robot

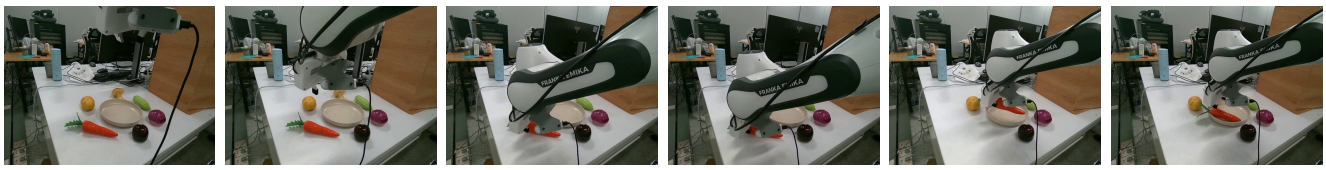
Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 2

[4] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2

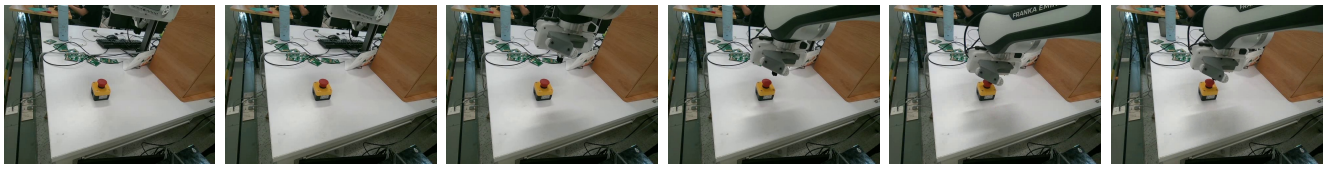
[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[6] Chuan Wen, Xingyu Lin, John Ian Reyes So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *Robotics: Science and Systems*, 2024. 1

Put the bitter guard on the plate.

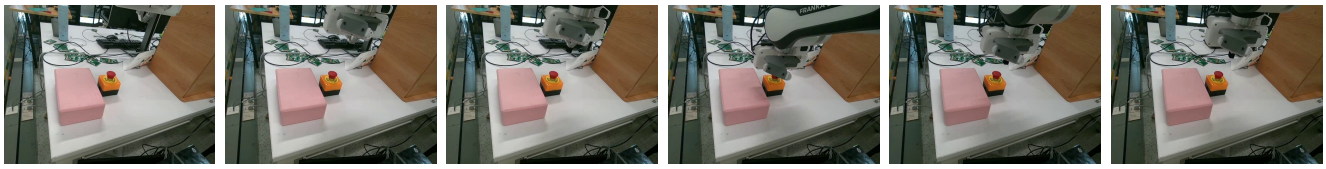


Put the wild carrot on the plate.



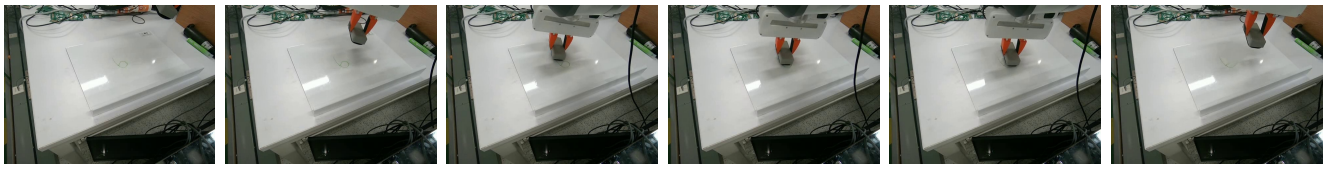
Press the button.

Press the button.



Press the button.

Wipe the white board.



Wipe the white board.

Figure 6. Real World Franka Robot.