A Appendix

A.1 Limitations

Our work prsesents a method enabling robots to process instructions and observations across various modalities, translating them into actionable movements. Despite conducting numerous experiments in both real-world and simulated environments, the method might not perform as expected in untested settings. We plan to release our datasets upon publication, but it's important to note that they were collected in a specific setting requiring a Franka Emika robot for manipulation tasks. Performance may vary in different environments due to factors such as lighting conditions.

Our methodology currently supports a variety of input modalities for commanding a robot to act under different circumstances, specifically tested on the Franka robot. We aim to broaden our experimental scope to include other robotic systems, such as the UR5, though it's important to note that our approach is not yet adaptable to different robotic embodiments. Furthermore, there are several unexplored modalities, including tactile feedback, sound-for-manipulation, infrared, and brain wave signals, that could significantly enhance human-robot interaction and improve model performance. Unfortunately, our method has not been evaluated with these modalities, primarily due to the limitations in acquiring the necessary equipment for data collection. Expanding our research to integrate these modalities represents a promising direction for future work, potentially unlocking new dimensions in robotics.

A.2 Social Impact

Our method facilitates the use of diverse input modalities to command robots to act under a variety of circumstances. However, it's important to acknowledge that increasing the number of input modalities also enlarges the attack surface for potential hackers. This expansion raises the risk of malicious parties jailbreaking the model, gaining control over the robot, and commanding it to perform unethical or harmful actions. It underscores the critical need for robust security measures to safeguard against such vulnerabilities.

A.3 More Experiments

Experimental results on Maniskill-2. We conducted a comparative analysis of our method against ResNet152 [87] and Swin Transformer Base [88], both of which are pre-trained on the ImageNet dataset. The baseline architecture follows the Maniskill-2 framework, utilizing PointNet [89] for processing point cloud data. This representation was then integrated with the image branch as delineated in Maniskill-2. In both experimental setups, our method consistently outperformed the two baseline models. In the dual-modal observation setting, our Any2Policy model achieved a significantly higher average success rate. It is also worth noting that, adding point cloud results in worse performance on some tasks for Swin Transformer. Yet, for our method, adding point clouds gives a significant boost to the success rate. These experimental results further support that our designed architecture can be beneficial to more modalities at the training stage.

Effectiveness of cross-attention in aligning instruction and observation. In Table 5, we compare the efficacy of cross-attention with FiLM [85], a method for aligning text with a visual backbone. Cross-attention demonstrates a markedly higher success rate than FiLM in all instruction-observation modality pairs. We think this can be caused by difficulty in optimizing modalities like point cloud. It highlights the effectiveness of the proposed cross-attention mechanism that uses external queries to support multiple modalities.

Table 5: Comparison between different instruction-observation alignment techniques.

Instruction → Observation	Text → Image	Text → Video	Text → Point Cloud	Audio → Image	Audio → Video	Audio → Point Cloud
Cross-Attention (Ours)	**51**	**57**	**62**	**49**	**55**	**57**
FiLM [85]	18	21	17	32	14	5

Instruction → Observation	Image → Image	Image → Video	Image → Point Cloud	Video → Image	Video → Video	Video → Point Cloud
Cross-Attention (Ours)	**56**	**63**	**38**	**46**	**57**	**36**
FiLM [85]	15	11	5	2	7	6

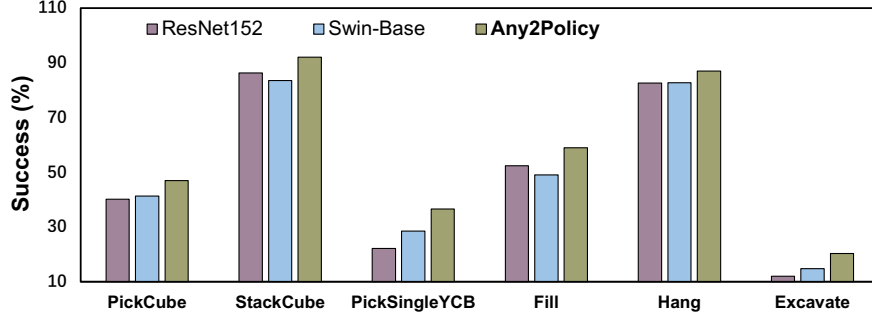
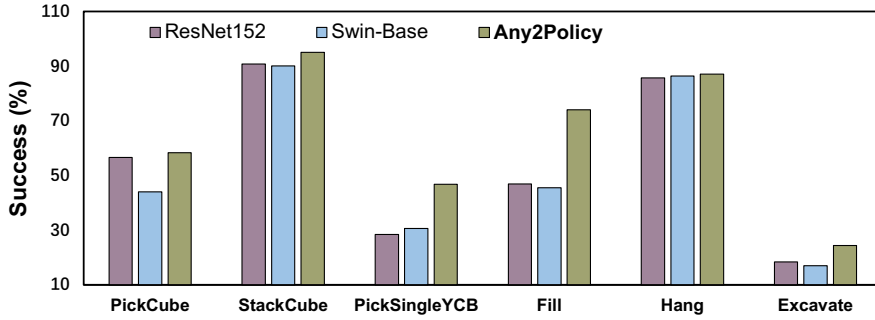
23

Performance comparison in Maniskill-2 with RGB

Performance comparison in Maniskill-2 with RGB + Point Cloud

Figure 5: Performance of Any2Policy in Maniskill-2. Adding point clouds (right) boosts the performance of Any2Policy.

A.3.1 Datasets

In our paper, we detail the simulated environments used for our experiments. The Franka Kitchen benchmark focuses on activities such as sliding the right door open, opening cabinets, turning on lights, adjusting stovetop knobs, and opening microwaves. Meta-world offers a set of complex tasks that test advanced object manipulation skills, including assembling a ring on a peg, picking and placing blocks between bins, pushing buttons, opening drawers, and hammering nails. ManiSkill2 features six tasks divided between three involving rigid bodies and three with soft bodies. These tasks are PickCube, StackCube, PickSingleYCB, Fill, Hang, and Excavate. For all tasks in both environments, proprioceptive data, including the positions of the robot's arm joints and gripper, are included.

A.3.2 More Training Details

In addition to the implementation specifics outlined in §4.1, our approach incorporates data augmentation techniques for images, encompassing adjustments in brightness, contrast, saturation, hue, and spatial translation. Due to the significant memory demands associated with handling multiple modalities, we have set the batch size to 16.

A.3.3 Token pruning.

In our embodied alignment model, we employ TokenLearner to significantly reduce the number of visual tokens: image tokens by 91%, video tokens by 96%, and point cloud tokens by 93%. This reduction is vital for both training and inference phases. During training, the involvement of multiple modalities would result in substantial GPU memory consumption if all tokens were

24

Make coffee

Put item inside the box if it has the same color as the box

Open microwave

Open the top drawer

Open the blue box and place the head near banana

Put cubic inside drawer

Put orange toy car inside drawer

Rearrange objects according to the end-goal image

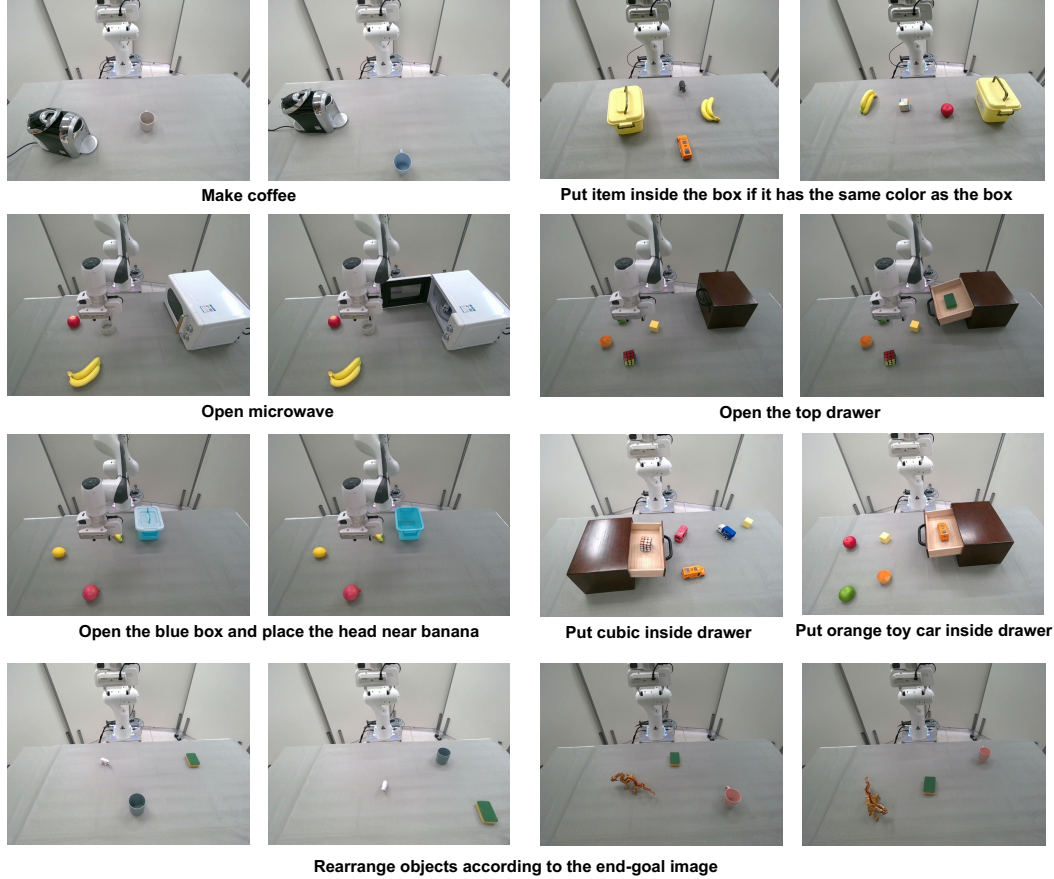



Figure 6: More qualitative examples of our collected datasets.

retained. Similarly, in the inference stage, using multiple modalities without token pruning would lead to considerable latency in action prediction. It's important to note that we cannot offer a direct performance or speed comparison with methods that do not utilize token pruning. The reason is that training with all tokens is infeasible, as it exceeds our GPU's memory capacity, preventing us from obtaining results using full token sets. Note that we reduce image token to 8 and video token to 16 because it is pre-defined by TokenLearner.

A.4 Qualitative Examples of Collected Data

In our study, we introduce a novel dataset comprising various modalities of task specifications and observations. Figure 3 showcases initial examples of the tasks included in our dataset, and herein, we provide additional examples to further illustrate the scope of our work. We have constructed multiple environments featuring the Franka robot, each with a slightly different setup with the same hardware. This diversity ensures that our approach is robust and can be generalized to new environments, demonstrating its adaptability and potential for widespread applicability in robotics research. The examples are presented in Figure 6.

A.4.1 Notes on Audio Modality

In our approach, language instructions are directly converted into audio signals, which are then utilized to command the robot, thereby eliminating the need for text conversion. This method is distinct from previous efforts that sought to improve robot model performance by employing sound-of-manipulation techniques [90, 91]. Such approaches often encounter challenges due to environmental noise and the need for extensive sensory data collection. Our method capitalizes on audio data as an immediate form of instruction, allowing the robot to execute tasks without the intermediate step

25

of text conversion. This strategy not only expedites the inference process but also introduces an additional method for human-robot interaction, underscoring its practicality in the field.

A.5 Video Demos

Our video demo can be found in https://any2policy.github.io/.

26