ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation

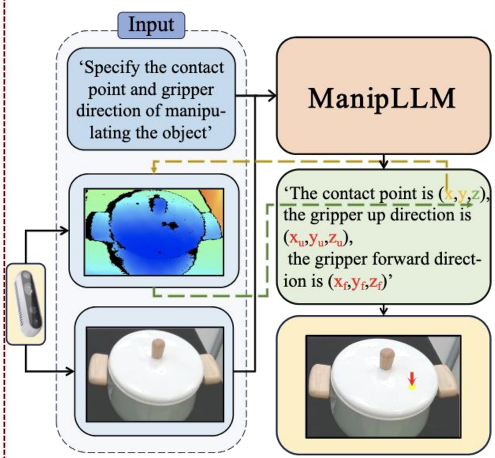
Xiaoqi Li[1], Mingxu Zhang[2], Yiran Geng[1], Haoran Geng[3], Yuxing Long[1],
Yan Shen[1], Renrui Zhang[3], Jiaming Liu[1], Hao Dong[1]
[1] Peking University [2]Beijing University of Posts and Telecommunications,
[3]The Chinese University of Hong Kong

1. Background

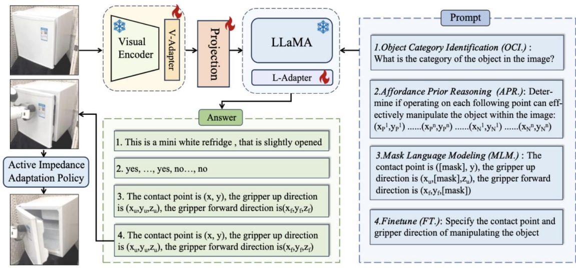
Robot manipulation relies on accurately predicting contact points and end-effector directions to ensure successful operation. However, learning-based robot manipulation, trained on a limited category within a simulator, often struggles to achieve generalizability, especially when confronted with extensive categories

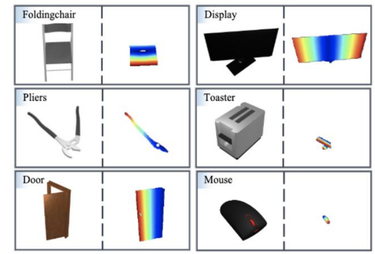
2. Limitations

We introduce an innovative approach for robot manipulation that leverages the robust reasoning capabilities of Multimodal Large Language Models (MLLMs) to enhance the stability and generalization of manipulation. By fine-tuning the injected adapters, we preserve the inherent common sense and reasoning ability of the MLLMs while equipping them with the ability for manipulation. The following figure shows the prediction process. Given the text prompt, RGB image, and depth map inputs, we obtain 3D pose of end effector.



3. Overall Framework





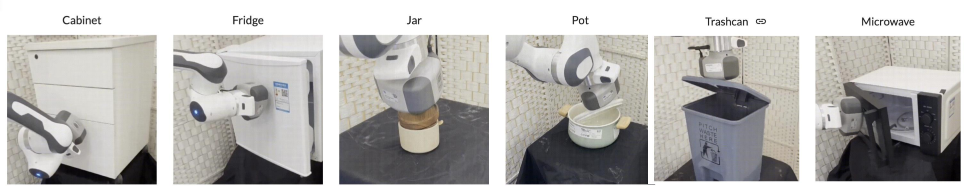
The left figure shows the overall training framework of four training tasks, enabling the model to recognize the current object (category-level), understand which regions can be manipulated (region-level), and finally generate a precise end-effector pose (pose-level). The right figure shows the ground truth of affordance prior.

4. Experiment

1) The table shows the results in simulator, in which ManipLLM shows strong performance across various category.
2) Meanwhile, the perforance in real-world is also stable across categories and view angles.

Method	Train Categories															
Where2Act [23]	0.26	0.36	0.19	0.27	0.23	0.11	0.15	0.47	0.14	0.24	0.13	0.12	0.56	0.68	0.07	0.40
UMPNet [33]	0.46	0.43	0.15	0.28	**0.54**	0.32	0.28	0.56	0.44	0.40	0.10	0.23	0.18	0.54	0.20	0.42
FlowBot3D [6]	0.67	0.55	0.20	0.32	0.27	0.31	0.61	**0.68**	0.15	0.28	0.36	0.18	0.21	0.70	0.18	0.26
Implicit3D [39]	0.53	0.58	0.35	0.55	0.28	**0.66**	0.58	0.51	0.52	0.57	0.45	0.34	0.41	0.54	0.39	0.43
Ours	**0.68**	**0.64**	**0.36**	**0.77**	0.43	0.62	**0.65**	0.61	**0.65**	**0.52**	**0.53**	**0.40**	**0.64**	**0.71**	**0.60**	**0.64**
Ours (long)	0.68	0.62	0.28	0.76	0.43	0.62	0.65	0.61	0.61	0.45	0.43	0.38	0.62	0.71	0.60	0.63

Method	Train Categories				AVG	Test Categories										AVG
Where2Act [23]	0.13	0.18	0.13	0.40	0.26	0.18	**0.35**	0.38	0.28	0.05	0.21	0.17	0.20	0.15	0.15	0.21
UMPNet [33]	0.22	0.33	0.26	0.64	0.35	0.42	0.20	0.35	0.42	0.29	0.20	0.26	0.28	0.25	0.15	0.28
FlowBot3D [6]	0.17	0.53	0.29	0.42	0.37	0.23	0.10	0.60	0.39	0.27	0.42	0.28	0.51	0.13	0.23	0.32
Implicit3D [39]	0.27	0.65	0.20	0.33	0.46	**0.45**	0.17	0.80	0.53	0.15	0.69	0.41	0.31	**0.30**	0.31	0.41
Ours	**0.41**	**0.75**	**0.44**	**0.67**	**0.56**	0.38	0.22	**0.81**	**0.86**	**0.38**	**0.85**	**0.42**	**0.83**	0.26	**0.38**	**0.51**
Ours (long)	0.37	0.75	0.44	0.67	0.54	0.34	0.22	0.81	0.86	0.30	0.85	0.42	0.80	0.26	0.38	0.50



Xiaoqi Li[1], Mingxu Zhang[2], Yiran Geng[1], Haoran Geng[3], Yuxing Long[1],
Yan Shen[1], Renrui Zhang[3], Jiaming Liu[1], Hao Dong[1]
[1] Peking University [2]Beijing University of Posts and Telecommunications,
[3]The Chinese University of Hong Kong