

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

Xiaoqi Li¹, Mingxu Zhang², Yiran Geng¹, Haoran Geng¹, Yuxing Long¹,
Yan Shen¹, Renrui Zhang³, Jiaming Liu¹, Hao Dong[†]

¹School of Computer Science, Peking University

² Beijing University of Posts and Telecommunications ³ MMLab, CUHK

A. Additional Experiment Details

A.1. More Details on Experiment Setting

When collecting training data, to augment domain randomization, we place the camera 4.5-5.5 units away from the object, facing the object’s center, and situated in the upper hemisphere of the object at a random azimuth angle between 0° and 360° , as well as a random altitude angle between 30° and 60° . This boosts the variety in view angle and help to deal with view angle issue when transferring from simulator to real world.

In simulator, we’ve also employed domain randomization to amplify scenario diversity, diversifying elements like lighting, materials, light position, etc, aiming to ease sim-to-real transfer. We visualize the domain randomization of handle material in Fig. 1.

In order to tackle the significant disparity between visual and collision shapes, we leverage the V-HACD [2](Voxelized Hierarchical Approximate Convex Decomposition) algorithm. This method entails voxelizing the 3D model, subsequently engaging hierarchical approximation to iteratively diminish the voxel count and amalgamate them into larger convex voxels. Subsequently, convex decomposition is applied to transform these merged convex voxels into simpler convex shapes.

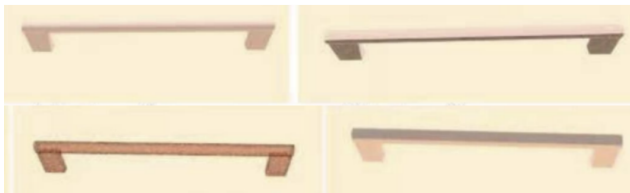


Figure 1. Domain randomization on material.

A.2. Representation for Each Category Icon

In Table 1, we provide an overview of the meaning of each category icon in Table 1 in the main paper. These categories, along with their corresponding objects, are sourced from PartNet-Mobility [1].

B. More Experiments

B.1. Experiments for TTA in Simulator

Because TTA (Test-Time Augmentation) is a plug-and-play strategy, we also employ this approach in the simulator when facing test categories to analyze deeper into its effectiveness. In this experiment, we utilize the success or failure of manipulations in the simulator as a supervisory signal to guide the model in determining whether the predicted pose will lead to a successful manipulation outcome. We only update the visual encoder’s V-Adapter to preserve the model’s inherent capabilities as much as possible while adapting to the target domain. Under this testing strategy, for the measurement of the initial movement in the test category, the success rate increases from 0.51 to 0.54, indicating an improvement with this strategy. Moreover, to maintain the model’s generalization performance, the number of updated parameters is minimal. The model’s capacity for updatable parameters is not extensive, resulting in a moderate increase in the success rate, showing it is still the model’s intrinsic capabilities playing a more dominant role.

We further investigate and find that when statistically testing the initial 50 test samples, the manipulation success rate increases significantly, showing an improvement of approximately 0.05 compared to the same period without TTA. In subsequent tests, the rate of improvement slows down. In the final 50 test samples, the improvement is approximately 0.01 compared to the same period without TTA. We thus assume that due to the limited number of parameters in the V-Adapter, there is a finite amount of knowledge that can be learned, and the potential for performance improvement is not limitless.

[†]Corresponding author: hao.dong@pku.edu.cn































									
Safe	Door	Display	Fridge	Laptop	Lighter	Microwave	Mouse	Box	Trashcan
									
Pot	Suitcase	Pliers	Storage	Remote	Bottle	Foldingchair	Toaster	Lamp	Dispenser
									
Toilet	Scissors	Table	Stapler	Kettle	USB	Oven	Washingmachine	Faucet	Phone

Table 1. Representation of each category icon.

To verify this, we add adapters to more layers in the visual encoder. Our approach (in main paper) involves adding adapters only to the linear layers in the Clip encoder. In the comparative experiment, adapters are added to the transformer layers as well, increasing the number of learnable parameters more than ten times. In this scenario, the total manipulation rate remains comparable to the test without TTA (0.51). In the initial 50 test samples, the increased number of learnable parameters quickly improves the model’s performance in the target domain by 0.07. However, in subsequent stages, the model’s performance even lags behind the test strategy without TTA. This indicates that allowing more model parameters to adapt to the target domain may result in a loss of the model’s original generalization. Therefore, we come to the conclusion that there is a trade-off between the size of learnable parameter and manipulation performance.

B.2. Quantitative Results in Simulator

In Fig. 2, we visualize the initial and final object state to demonstrate how does the robot manipulate in the simulator. The distinction becomes more apparent when zoomed in at a factor of four.

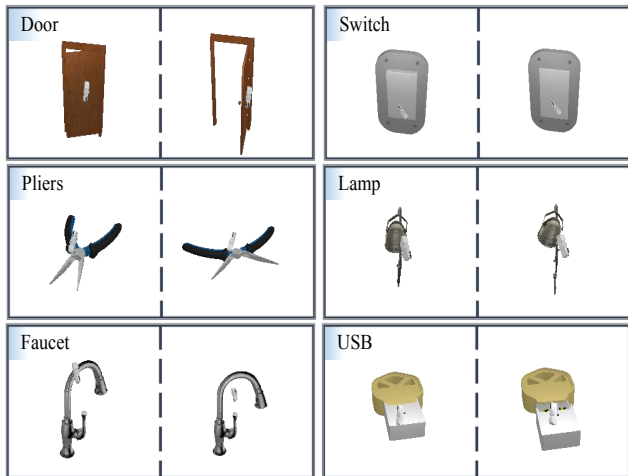


Figure 2. Manipulation demonstration in simulator.



(a) trash can



(b) fridge

Figure 3. Manipulation demonstration in real-world.

C. Real World Experiments

In this section, we analysis the limitation and failure cases that in our real-world setting. We observe that the primary limitation still lies in the potential for the suction cup to collide with the object’s surface, especially if its orientation is not adjusted appropriately. Additionally, there is a possibility that the suction cup may fail to hold the object, as it requires a specific pressure between the cup and the object to establish a vacuum and effectively hold the item in place. Video demonstrations are shown in the **supplementary video**. In Fig. 3, we present snapshots of partial real-world experiments, illustrating the initial object state, initial contact state, and final contact state, respectively.

References

- [1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1
- [2] Khaled Mamou, E Lengyel, and A Peters. Volumetric hierarchical approximate convex decomposition. In *Game Engine Gems 3*, pages 141–158. AK Peters, 2016. 1