

Supplementary Material of Autonomous Interactive Correction MLLM for Robust Robotic Manipulation

Anonymous Author(s)

Affiliation

Address

email

1 More related works.

Robotic Manipulation. Reinforcement learning[39,40,41,42], imitation learning[43,44,45], and deep learning for visual understanding[46,33,47,32] have been extensively applied in robotic manipulation. Recent studies like where2act[46] and Flowbot3d[33] combine prior knowledge with deep learning. Where2act[46] leverages point cloud data as input to initially predict the score of each point, followed by rotation prediction. UMPNet[32] employs a position network for position inference, then utilizes a action sampler to sample some directions of this position, and subsequently employs two networks to evaluate these directions. Flowbot3D[33] introduces an articulation flow to represent the point-wise potential motion and uses this representation to guide the manipulation. Flowbot++[47] integrates joint axis modeling into Flowbot3D to make the trajectory more smooth. Following the emergence of powerful large language models, researchers have started employing them to address the generalization challenges in manipulation tasks. Our work also focuses on leveraging large-scale models to enhance the robustness of algorithm in manipulation task.

Multimodal Large Language Models(MLLM). Large language models(LLM) have impressive power in natural language processing tasks. Consequently, some researchers aim to develop even more powerful MLLMs based on LLMs to tackle a broader range of vision-language tasks[48,49,9,37,36,50]. CLIP[37] first established a connection between language and vision. The BLIP series[48,49,9] freezes the LLM and the image encoder, training their bridge, Q-Former, to align the two modalities. It also employs instruction tuning to enhance its ability to follow instructions. LLaVA[50] use a simple fully connected layer to build the bridge of the LLM and the image encoder. LLaMa-Adapter[36] utilizes projection and adapters to make the model have the multimodal ability and reduce training costs. Additionally, closed-source MLLMs like GPT-4V[51] are more powerful than the open-source models we mentioned previously.

2 Detailed about Methodology and Implementation.

2.1 Rotation-related Information Extraction

In this section, we will provide a detailed introduction to how we perform joint type classification. After the end effector (EE) reaches the predicted pose and adheres to the object, it will move along the predicted gripper direction for a certain distance. We will record the pose trajectory of EE for 20 frames during this process. We denoted the i -th pose in the trajectory as ${}^wP_i^E$. We perform the operation of subtracting the previous pose from the current pose for all poses in the sequence which can be denoted as $\vec{v} = {}^wP_i^E - {}^wP_{i-1}^E$. We gather the \vec{v} to form a vector list. Then we calculate the angle between adjacent vectors. Next, we set a threshold. If all the angles are less than this threshold, then the joint will be considered a prismatic joint; otherwise, it will be considered a revolute joint.
















2.2 More Implementation Details














In this section, we will include additional implementation details that were not mentioned in the main paper. During the training phase, for the VQA task 'Mask Position Reasoning,' we mentioned that we randomly select N points. In implementation, N is set to 20. During testing, it is important to note that all MLLM temperatures are set to 0. In the tta training process, we initialize the learning rate(lr) to $5e-8$ and set the weight decay to $2e-3$. Every 300 iterations, we reduce the lr by 70%.

3 More Ablation

Tab. 1 presents the ablation study results comparing the Vip-llava method and the GT method. Vip-llava uses the Vip-llava model to predict part mobility, while GT uses the ground truth interaction map. Across various categories, Vip-llava has some performance loss compared to GT, with average scores of 0.69 and 0.68 for training and test categories, respectively, compared to GT's higher averages of 0.76 and 0.76. This indicates that Vip-llava's understanding of embodied data is insufficient. Utilizing models with better comprehension of embodied data or incorporating human feedback into our framework could lead to improved results.

Table 1: Ablation study

Method	Train Categories														
															
Vip-llava	0.86	0.73	0.50	0.25	0.81	0.80	0.68	0.62	0.38	0.68	1.00	0.86	0.55	0.83	0.38
GT	0.81	0.78	0.75	0.78	0.81	0.67	0.86	0.85	0.44	0.79	1.00	0.75	0.74	0.83	0.73

Method	Train Categories					Test Categories									
					AVG										AVG
Vip-llava	0.60	0.70	0.27	1.00	0.69	0.36	0.45	0.76	0.83	0.64	0.86	0.70	0.78	0.64	0.68
GT	0.40	0.80	0.66	1.00	0.76	0.55	0.79	0.77	0.33	0.36	0.92	0.43	0.71	0.79	0.76

4 Example of Prompts.

4.1 Position correction prompt

Firstly, we make use of the mask information to help the MLLM get a more precision prediction by the following prompts.

Is there a red mask in the image?
AIC-MLLM: No.
Is there a red mask in the image?
AIC-MLLM: Yes.

Determine if following point is in the red mask: (172,47),(117,192),(323,251),(195,9),(211,277),(242,292),(87,70),(88,314),(193,39),(87,174),(88,165),(25,333),(72,265),(115,243),(197,335),(99,177),(243,285),(147,147),(288,265),(185,127)
AIC-MLLM: No,Yes,No,No,No,No,No,No,No,No,No,No,No,No,No,Yes,No,Yes
Determine if following point is in the red mask: (11,104),(212,138),(182,125),(156,111),(258,27),(217,151),(309,307),(174,148),(29,67),(35,295),(73,297),(302,218),(259,287),(265,27),(199,61),(44,290),(88,33),(133,232),(255,36),(256,290)
AIC-MLLM: No,No,Yes,Yes,Yes,No,No,Yes,No,No,No,No,No,No,No,No,No,No,No,No,No,No,No,No

For position correction using the red mask image, we reassess the validity of the previously predicted contact point. If the contact point is deemed ideal, the MLLM will confirm this by replying "Yes." Otherwise, it will reply "No."

Given the part with a red mask known to be unmovable in the image, is the (112,171) considered as the successful contact point?
AIC-MLLM: Yes.
Given the part with a red mask known to be unmovable in the image, is the (160,160) considered as the successful contact point?
AIC-MLLM: No.

Then, MLLM makes the prediction using the image with red mask.

Specify the successful contact point and orientation of pulling the object. DONOT predict the successful contact point in the red mask part.
AIC-MLLM: The contact point is at (170, 179), the gripper direction is [48, -1, -12], the gripper forward direction is [-1, 1, 49].
Specify the successful contact point and orientation of pulling the object. DONOT predict the successful contact point in the red mask part.
AIC-MLLM: The contact point is at (143, 174), the gripper direction is [48, 1, 12], the gripper forward direction is [1, -49, -10].

MLLM assesses whether the predicted point is a successful point.

You pull the object in the picture by point(170,162),gripper direction[47,11,-11],gripper forward direction[10,-49,1],if it will success answer only Yes , else answer only No.
AIC-MLLM: No.
You pull the object in the picture by point(179,179),gripper direction[38,-28,13],gripper forward direction[13,45,-19],if it will success answer only Yes , else answer only No.
AIC-MLLM: Yes.

53 4.2 Rotation correction prompt

54 The FIE acquires environmental information, including joint type, joint axis and the normal direction
55 of previously predicted contact point. The MLLM uses this information, along with the contact point, to infer whether to correct its response by suggesting a new gripper direction.

You're trying to pull an object part, the joint's type is REVOLUTE, the joint axis direction is [-22, -1, 44], Your contact point is (79, 170), the norm direction of the point is [-47, 9, 12]. Please tell me your gripper direction.
AIC-MLLM: The gripper direction is [-47, 13, -13].
You're trying to pull an object part, the joint's type is REVOLUTE, the joint axis direction is [10, 6, -48], Your contact point is (170, 159), the norm direction of the point is [-29, 0, 40]. Please tell me your gripper direction.
AIC-MLLM: The gripper direction is [-28, 35, -20].

56 5 Example of Failed Case and correction process. 57

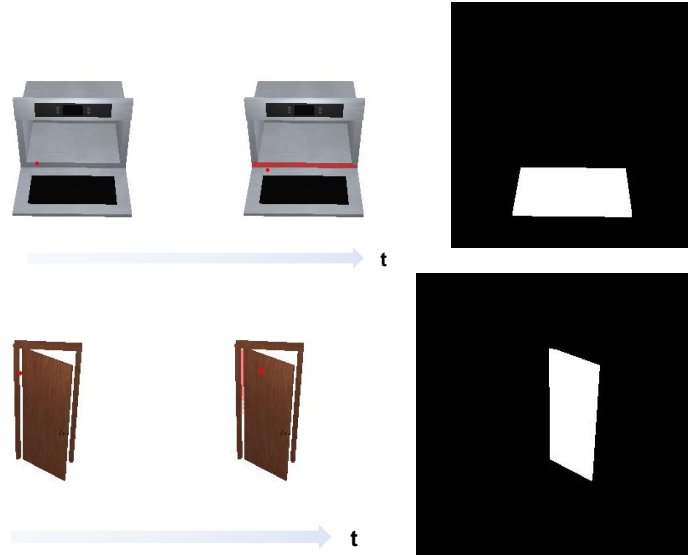


Figure 1: **Examples of position correction.** The left two figures are the prediction process of keeping away from the red mask. The third figure is the interaction map where white area is the movable part.



Figure 2: **Rotation correction.** Correction times increase sequentially along the timeline.

58 6 Real-world experiments

59 To validate our end-to-end method beyond simulations, we conducted real-world experiments using
60 a Franka Emika robotic arm equipped with an Intel RealSense D415 sensor. For more details, please
61 refer to the video and our website (<https://sites.google.com/view/aic-ml1m>).

62 **Experimental Setup.** The setup included a Franka Emika Panda robotic arm, known for its preci-
63 sion and versatility. An Intel RealSense D415 sensor was used to capture RGB-D data, providing

64 3D information necessary for the robotic arm’s operations. We closed the traditional gripper and
65 applied double-sided tape to its head as a suction gripper.

66 **Initialization.** The robotic arm and D415 sensor were calibrated to ensure accurate spatial data
67 capture and precise movements.

68 **Image Processing.** The D415 sensor captures high-resolution RGB-D images, which are processed
69 to meet the requirements of our pipeline. Specifically, our method necessitates images sized at
70 336x336 pixels. To preserve information during resizing, we apply both cropping and padding
71 techniques. This involves cropping or adding white borders around the captured images to resize
72 them to 336x336 pixels. The resulting image, whether cropped or padded, is then used for further
73 analysis in our end-to-end pipeline.

74 **Task Execution.** Using our end-to-end method, the robotic arm was directly controlled to perform
75 tasks based on the RGB-D data input. The tasks involved the arm autonomously approaching and
76 manipulating objects within the workspace.

77 **Demonstration.** The robotic arm demonstrated its ability to perform these tasks seamlessly, show-
78 casing the practical applicability and robustness of our end-to-end method in handling real-world
79 scenarios.