AR-VRM: Imitating Human Motions for Visual Robot Manipulation with Analogical Reasoning
Supplementary Material

Dejie Yang[1], Zijing Zhao[1], Yang Liu[1,2*]
[1] Wangxuan Institute of Computer Technology, Peking University
[2] State Key Laboratory of General Artificial Intelligence, Peking University
{ydj,zijingzhao}@stu.pku.edu.cn, yangliu@pku.edu.cn

In this supplementary material, we provide a more detailed illustration and evaluation of our proposed AR-VRM. We present the details of real robot experiments in Section 1 to demonstrate the generalization capabilities of our AR-VRM. In Section 2, we present additional ablation studies and analysis, including the impact of various video retrieval approaches, the effectiveness of the number of retrieved videos, the effectiveness of different types of keypoint usage, and the inference time of DP and RAP. In Section 3, we outline further implementation details of AR-VRM, covering aspects such as the pretraining and fine-tuning datasets, as well as other experimental settings. In Section 4, we showcase additional qualitative examples.

1. Detials of Real Robot Experiment

To validate AR-VRM's effectiveness in real-world applications, we implemented physical experiments focusing on two key tasks: object transportation and articulated object manipulation.

Object Transportation Our experimental setup involved a straightforward arrangement consisting of a plate containing three objects: an orange, an apple, and a green jujube. We compiled a total of 1200 demonstrations, with each demonstration illustrating the movement of a single object. We evaluated the performance of AR-VRM across three distinct experimental settings: **1. Seen Objects.** In this setting, the robot was tasked with transporting the three objects observed during training: an orange, an apple, and a green jujube. In addition to evaluating the robot in a scene identical to the training environment (containing only these three objects), we further assessed its robustness in two unseen, disturbance-prone scenes. In the first disturbed scene, distractor objects — a tomato, a corn, and a yellow peach — were added alongside the original three objects. In the second disturbed scene, the background was altered by introducing a wooden board and a bowl. These additional tests allowed us to measure AR-VRM's ability to handle environmental disturbances effectively. **2. Unseen Instances.** This setting aimed to evaluate AR-VRM's generalization to unseen instances of the trained object categories. Specifically, the robot was instructed to transport a novel set of an orange, an apple, and a green jujube, all of which were different instances from those in the training data. **3. Unseen Categories.** Finally, to assess AR-VRM's capacity to generalize to entirely new object categories, the robot was tasked with transporting objects belonging to unseen classes during training: a tomato and a yellow peach.

Articulated Manipulation This experiment focused on evaluating AR-VRM's performance in contact-rich articulated object manipulation. To this end, we selected a drawer as the target articulated object and collected 1,400 trajectories involving both opening and closing the drawer for training purposes. AR-VRM outperformed the two baseline methods by a significant margin. However, the model exhibited two typical failure modes: (1) failing to completely close the drawer during the closing task, and (2) failing to engage with the drawer handle when attempting to pull it open during the opening task. Despite these challenges, AR-VRM demonstrated superior robustness and reliability in articulated manipulation tasks compared to the baseline approaches.

2. Additional Ablations Studies

In this section, we present additional ablation studies and analysis to evaluate the effectiveness of different video retrieval approaches, the number of retrieved videos and the impact of different types of keypoint usage. All the experiments are conducted on CALVIN[4] dataset, and evaluated on NVIDIA A800 GPU * 8.

[*] Corresponding author

Retrieval Approach	Success rate of tasks completed in a row					Avg.Len.
	1	2	3	4	5	
No-Retrieval	0.892	0.871	0.810	0.781	0.710	4.06
Video Feature Retrieval	0.927	0.882	0.823	0.791	0.727	4.15
Text Feature Retrieval	0.940	0.901	0.830	0.800	0.732	4.20
Video+Text Feature Retrieval	**0.951**	**0.915**	**0.855**	**0.800**	**0.751**	**4.27**

Table 1. **Ablation study on the design choices of video retrieval approaches.**

Num. of Videos	Success rate of tasks completed in a row					Avg.Len.	Inference Time(s) per sample↓
	1	2	3	4	5		
0	0.892	0.871	0.810	0.781	0.710	4.06	2.07
1	0.912	0.879	0.820	0.789	0.723	4.12	2.31
5	0.933	0.891	0.827	0.795	0.733	4.18	3.02
10	**0.951**	**0.915**	**0.855**	**0.800**	**0.751**	**4.27**	3.54

Table 2. **Ablation study on the number of the top J retrieved videos.**

Effectiveness of different video retrieval approaches. In Table 1, we show the impact of different video retrieval approaches on the performance of success rate of tasks completed. The following conclusions can be drawn: (1) Row 1 *No-Retrieval* denotes only fine-tuning and reasoning with robot data. Compared with robot-data-only fine-tuning, incorporating retrieved videos (no matter in what way) during the fine-tuning and reasoning process continually leads to an improvement in success rates. Notably, when using *Video+Text Retrieval* (row4), optimal performance is achieved(a gain of +0.21 on Avg. Len.). (2) Compared to using video retrieval alone (row2, *Video Feature Retrieval*), relying solely on text retrieval (row3, *Text Feature Retrieval*) brings greater benefits (+0.05 on Avg. Len.). This may be because videos (especially ego-centric videos) often contain various distractions such as objects, scene changes caused by camera shake which lead to inaccurate retrieval. Moreover, the visual context similarity may not always represents the relevance on actions of tasks. (3) Using both video features and text features for retrieval achieves the best results (row4). This may be because combining text and video features ensures that the retrieved videos are more relevant.

Influence of the number retrieved videos. In Table 2, we study the impact of the number retrieved videos. It can be observed that as the number of retrieved human videos increases, the success rate continues to improve (from 4.06 to 4.27 on Avg. Len. with the number of videos from 0 to 20). This indicates that human videos can provide guidance for the current manipulation task. With more videos providing more diversified scenes with richer semantic information, the robustness of guidance for robot arms is strengthened, thus resulting in better performance.

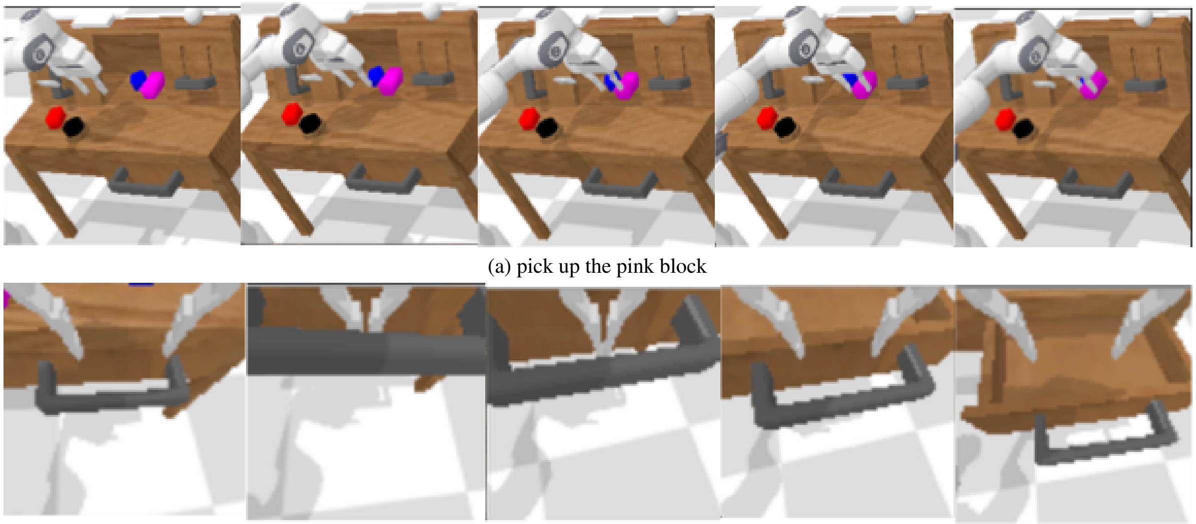
Effectiveness of different types of human keypoint usage. Using different types of human keypoints for the guidance of robot actions may results in different effects and performance. In Table 3, we compared three methods: not using keypoints (row1, our baseline which directly predicts frames), using full-body keypoints (row2, *Body*, using [5] to detect full-body keypoints including hands), and using hand keypoints (row3, *Hand*). (1) Compared to not using keypoints, incorporating body or hand keypoints brings varying degrees of performance improvement with different gains respectively . This indicates that keypoints, compared to video frames, focus on more critical information for action prediction, reducing the interference of background and other irrelevant information, enabling the model to concentrate on predicting key parameters and aligning better with the robotic arm's prediction and operation processes. (2) The performance of using hand keypoints is better than using full-body keypoints. Only under the evaluation of completing 4 and 5 tasks, full-body keypoints outperform hand keypoints by a small margin. This may be because most egocentric videos only contain hand not full-body. To achieve optimal performance, it may be more effective to focus on keypoints that are specifically aligned with the robotic arm's operational requirements, such as hand or arm-related keypoints, rather than using all full-body keypoints. In the future, we will explore introducing fine-grained human keypoints such as human arms, hands and whole-bodyxxxw, or investigate the relevance of different types of human keypoints for different types of robotic arms.

3. Additional Implementation Details

Datasets. We use the Ego4D dataset [2] for pretraining and the CALVIN dataset [4] for fine-tuning. Following

Keypoints	Success rate of tasks completed in a row					Avg.Len.
	1	2	3	4	5	
None	0.836	0.781	0.678	0.521	0.499	3.32
Human	0.929	0.904	0.839	**0.823**	**0.767**	4.26
Hand	**0.951**	**0.915**	**0.855**	0.800	0.751	**4.27**

Table 3. **Ablation study on the design choices of keypoints.**



(a) pick up the pink block

(b) open the drawer

Figure 1. **Examples of action predictions.**

[6], we leverage 20k expert trajectories from CALVIN, each paired with corresponding language instruction labels. These trajectories comprise delta XYZ positions and delta Euler angles for arm actions, as well as binary gripper actions. For evaluation, we utilize 1,000 unique sequential instruction chains. In each sequence, the robot aims to sequentially solve up to five tasks by interpreting a series of five language instructions. The evaluation is conducted across four distinct environments (A, B, C, and D), differentiated by desk colors and object configurations.

Other Experimental Settings. To enable efficient training, we pre-extract video frame features from both datasets using the Masked Autoencoder (MAE) [3]. For unseen instruction language generalization, we adopt GPT-4 [1] to generate 50 synonymous instructions per task. This is achieved using the prompt: "`Generate a synonymous sentence with new words for [task instruction]`." We follow the approach in [6], where these synonymous instructions are randomly sampled during evaluation. Here, `[task instruction]` refers to the language instructions provided in CALVIN [4].

4. Qualitative Results

Qualitative Analysis In Figure 1a and Figure 1b, we illustrate the action predictions for the tasks "pick up the pink block" and "open the drawer." The results demonstrate that the robot successfully interprets the language instructions and performs the corresponding actions. Specifically, for the task "pick up the pink block," the robot accurately identifies and picks up the correct object, showcasing its ability to associate the language instruction with the appropriate visual cues. Similarly, for the task "open the drawer," the robot effectively follows the instruction to locate and manipulate the specified object. These results highlight the model's generalization capability of object recognition and action understanding.

Failure Case and Analysis Figure 2 illustrates an example of an error in "in the drawer, grasp the blue block." The model incorrectly picked up the blue block from the cabinet instead of from the drawer. While the model successfully identified the target object, "blue block," it failed to understand the operational context, "drawer." Instead, it took a "shortcut" by directly grabbing the blue block from the cabinet. This issue may be related to the model's lack of

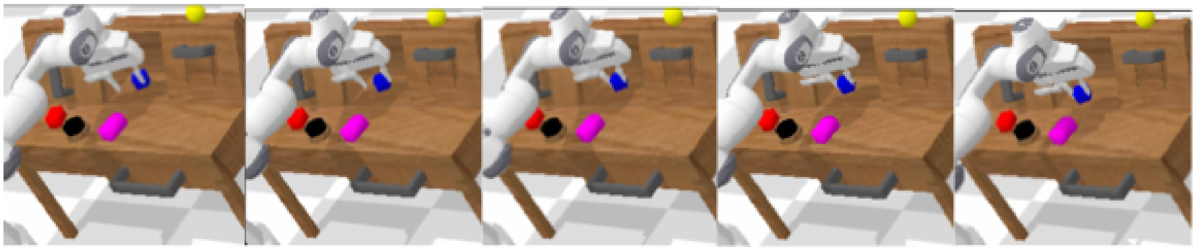


Figure 2. **A failure case**: in the drawer, grasp the blue block

consideration for multi-step operations (e.g., first locating the drawer, then locating the blue block). In future work, we will focus on designing mechanisms to better handle multi-step operations.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[4] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. 1, 2, 3

[5] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13031–13040, 2023. 2

[6] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 3