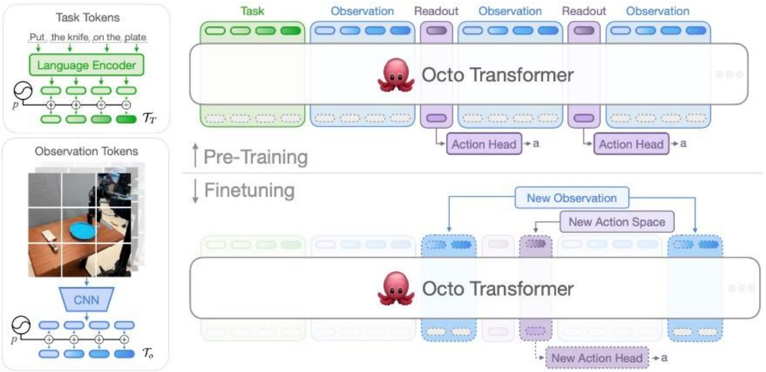
PIVOT-R: Primitive-Driven Waypoint-Aware World Model for Robotic Manipulation

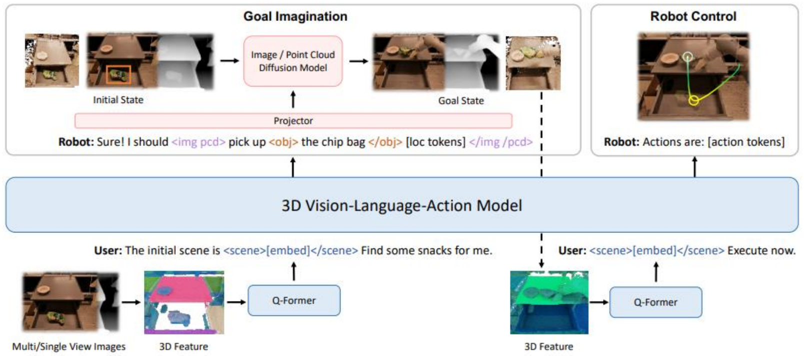
Kaidong Zhang*, Pengzhen Ren*, Bingqian Lin, Junfan Lin,
Shikui Ma, Hang Xu, Xiaodan Liang

https://abliao.github.io/PIVOT-R/

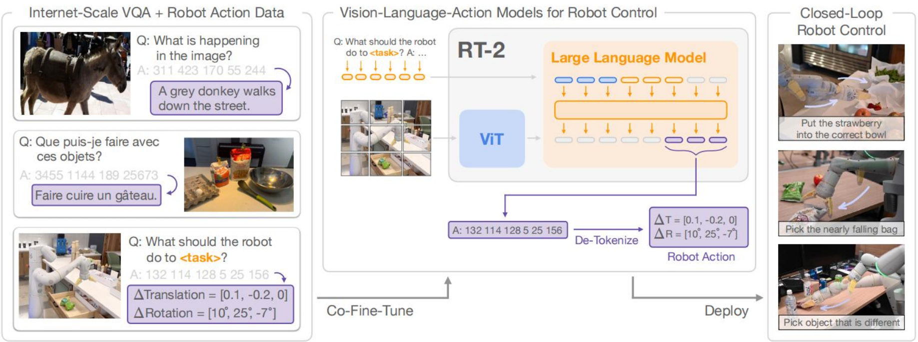
Motivation



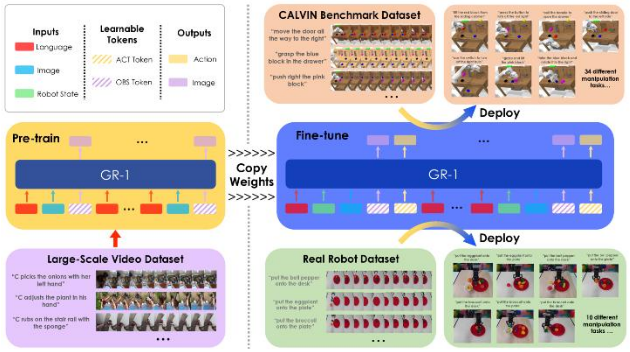
Octo: An Open-Source Generalist Robot Policy (2024)



3D-VLA: A 3D Vision-Language-Action Generative World Model



RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

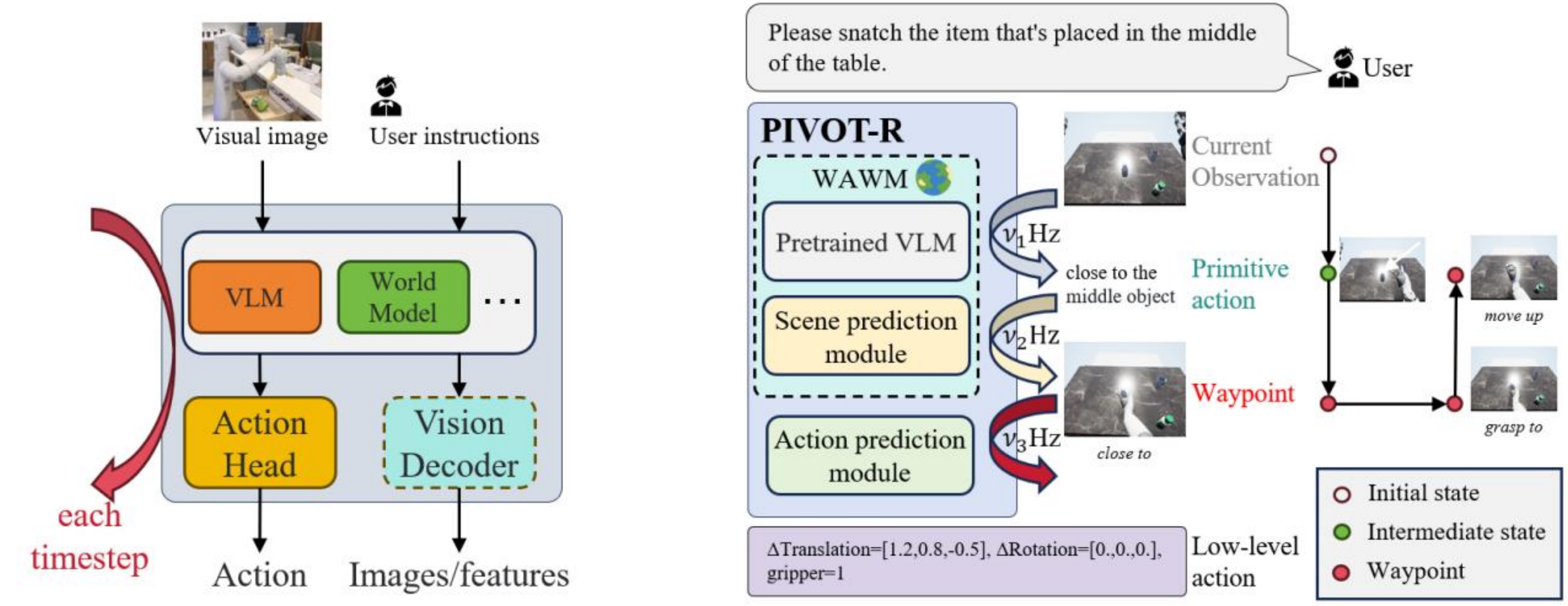


GR-1: Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation (2024)

- Fitting data to memorize the surficial pattern and thus are fragile to dynamic environment changes
- Larger models lead to lower efficiency

develop **efficient** **architecture** that can model **critical relationship between instruction and control signals**

Comparison



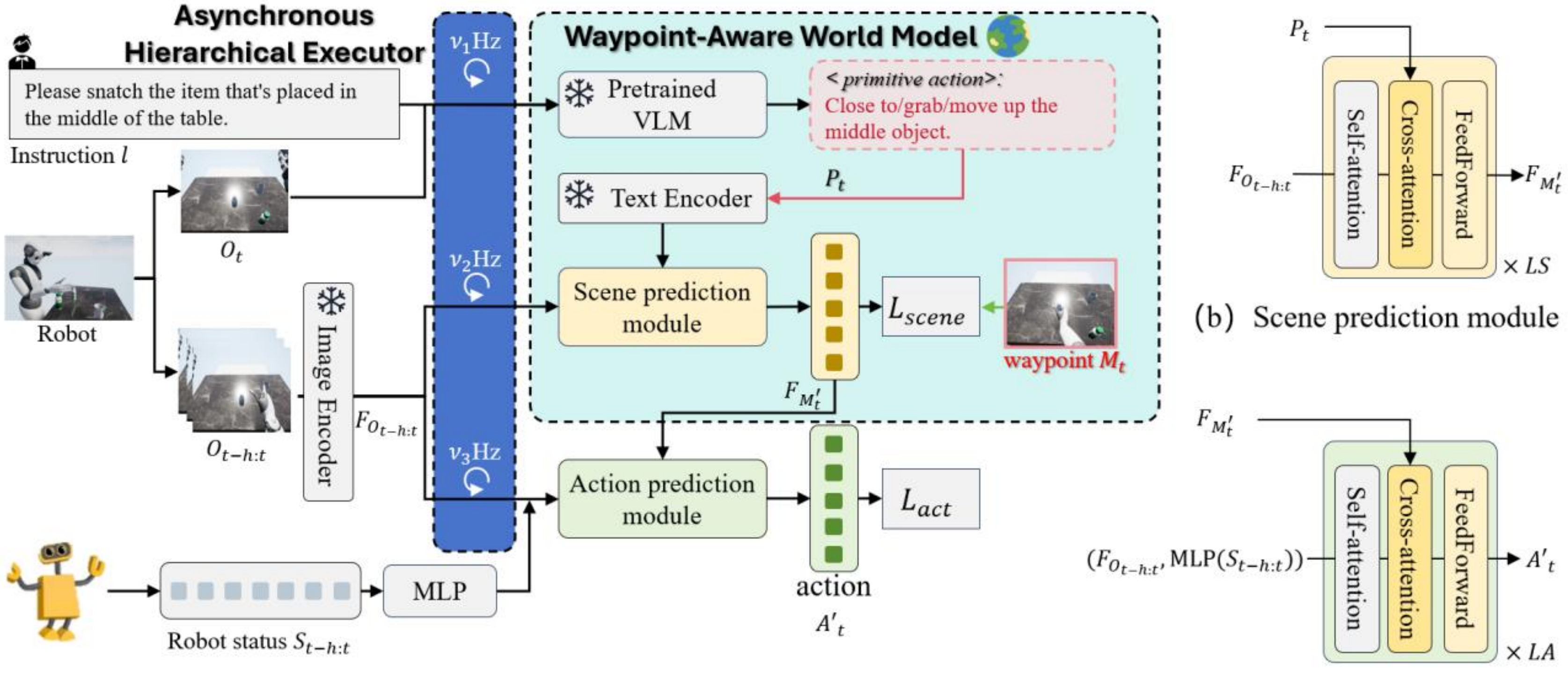
(a) Sequentially executed robot manipulation model

(b) PIVOT-R

Previous works:
- ➢ Fitting data through an end-to-end model and thus are fragile to dynamic environment changes
- ➢ Sequentially execute each module at each timestep, which leads to model redundancy

PIVOT-R:
- ➢ Focus on the prediction of waypoints related to the manipulation task, and acquire the transferable knowledge
- ➢ Parallel execution of each module to have higher execution efficiency

Overall Architecture



(a) Overall framework

(b) Scene prediction module

(c) Action prediction module

O_t: Observation image at time step t
S_t: Observation image at time step t
l: User's language instruction
F_O: Feature of Image
v_i: Frequency of the i-th module
P_t: Output of VLM
M_t: Waypoint at time step t
L: MSE

PIVOT-R consists of a waypoint-aware world model and a lightweight action prediction module.

- Waypoint-Aware World Model (WAWM). WAWM mainly includes a powerful VLM and a scene prediction module. VLM parses l to provide task-related waypoint prompts, which are used for guiding the scene prediction module to conduct critical waypoint prediction.
- Action Prediction Module. Receive sensors input and output of scene prediction module, and quickly output actions.

Results: Performance on Benchmark SeaWave

Model	Level 1	Level 2	Level 3	Level 4	Mean	Time(ms)
Gato	34.74	30.53	23.16	20.00	27.11	139
BC-Z	41.05	32.63	23.16	25.26	30.53	12
Octo	69.79	48.48	34.69	33.58	46.64	18
RT-1	67.38	49.47	38.95	34.74	47.64	21
GR-1	77.08	55.56	37.31	34.33	51.07	35
Surfer	74.74	61.05	45.26	37.89	54.74	24
PIVOT-R	**88.06** (13.32 ↑)	**77.55** (16.50 ↑)	**73.33** (28.07 ↑)	**57.82** (19.93 ↑)	**74.19** (19.45 ↑)	27

Success rate and speed comparison of different methods in four levels of tasks (%). PIVOT-R substantially achieved a significant improvement on all tasks. Specifically, PIVOT-R achieved an average success rate of 74.19%, 19.45% higher than the best baseline. Both the manipulation ability and the ability to understand instructions have been greatly improved. At the same time, the inference speed is not slowed down.

NEURAL INFORMATION PROCESSING SYSTEMS

Results: Performance on Benchmark SeaWave

Model	Seen	Unseen backgrounds	Changing lights	Distractors
Gato	24.56	20.83	23.33	16.67
BC-Z	27.02	19.17	18.33	21.67
Octo	38.92	40.83	37.50	35.83
RT-1	41.05	38.33	40.83	35.00
GR-1	42.40	40.83	35.00	37.50
Surfer	48.07	46.67	45.83	40.83
PIVOT-R	**69.57** (21.0 ↑)	**59.17** (12.5 ↑)	**61.67**(15.84 ↑)	**55.83**(15.0 ↑)

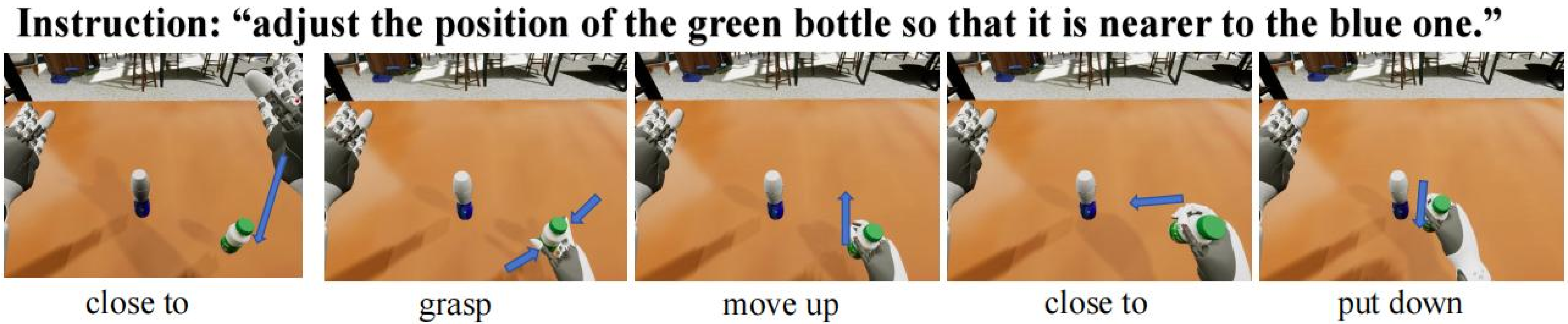
Performance comparison on seen scenarios, different backgrounds, changing lights, and more distractors (%). We also perform experiments in different unseen scenarios, including unseen backgrounds, changing light intensity, and more distractions. PIVOT-R still maintains a success rate far superior to other models, indicating that with the help of WAWM, the model captures key information and maintains good generalization in changing scenarios.

Results: Performance on Real-World

Model	Pick up	Put on	Push to	Mean
Octo	34.72	27.78	4.17	22.22
RT-1	40.28	22.22	19.44	27.31
GR-1	26.39	29.17	8.33	21.30
Surfer	41.67	29.17	**31.94**	34.26
PIVOT-R	**54.17**	**41.67**	25.00	**40.28**

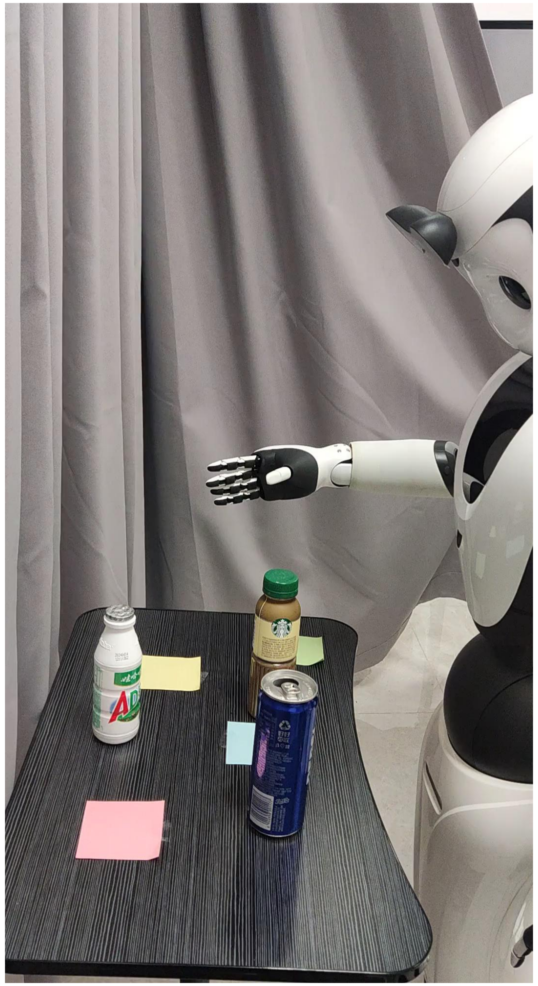
Performance of different methods on three real robot manipulation tasks (%). “Pick up”: pick up the correct object from the table. “Put on”: Pick up the object and place it on the correct color block. “Push to”: Push the object to the correct color block. PIVOT-R achieved the highest average success rate, with two tasks reaching best performance.

Qualitative results on SeaWave

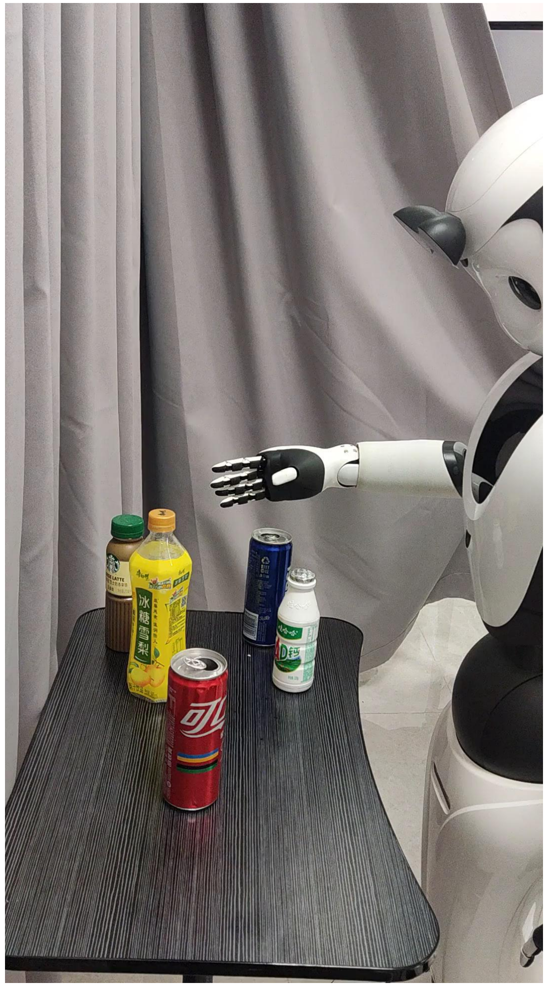


An example shows the execution process of PIVOT-R. It demonstrates the example of bringing milk close to yogurt. The task process can be divided into five actions. Through the instruction of primitive actions and the prediction of waypoints, the model successfully completes the task.

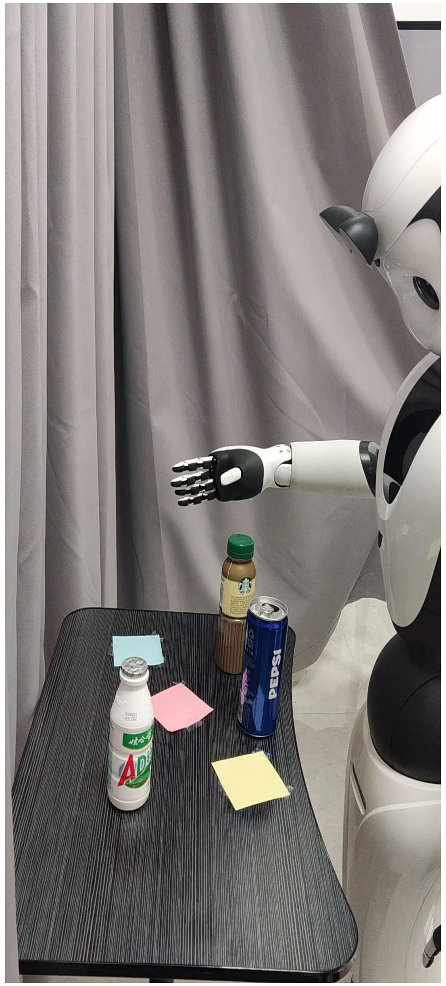
Visualization on Real-World



Pick the coffee and put on yellow block



Pick up the juice in the front row



Push coffee to pink block

Thank you!