

Annual Report

2025 Q4 ~ 2026 Q1



2025

Annual Report

Published by the RoboChallenge Committee

RoboChallenge Annual Report

2025 Q4 – 2026 Q1

With the explosive breakthroughs of large language models (LLMs) and visual language models (VLMs), artificial intelligence has demonstrated astonishing cognitive and reasoning abilities in the digital space. Today, this wave of technology is rapidly spilling over into the physical world, pushing robots to evolve from single-function automated devices to embodied intelligence (Embodied AI) with general understanding and decision-making capabilities. However, to bridge the gap from "lab-intelligence" to "real-world intelligence", the industry still needs to overcome challenges such as the difficulty of validation in real environments, non-standard testing conditions, and the lack of open platforms. Building an open, fair, and reproducible 'real-world testing ground' has become an essential path for the industry towards generalization.

RoboChallenge was born to address this core issue. As the world's first large-scale real machine evaluation platform for embodied intelligence, jointly initiated by Dexmal and Hugging Face, RoboChallenge is committed to providing more reliable and comparable evaluation benchmarks for visual-language-action models (VLAs) in practical applications of robotics through a scientific evaluation system. Since its launch on October 15, 2025, the platform has successfully deployed a testing cluster of 20 real machines, enabled with models such as UR5, Franka, ARX5, and ALOHA, and has open-sourced the Table30 dataset covering 30 standardized desktop tasks across 9 categories.

To further consolidate industry efforts and accelerate the standardization of real-world robot benchmark evaluation for embodied intelligence, the RoboChallenge Committee was officially established on November 20, 2025. Initiated by Dexmal and Hugging Face, the committee brings together a broad coalition of partners, including BAAI, Agibot, Qwen, Galaxea, X Square Robot, Tsinghua University, Xi'an Jiaotong University, and GOSIM, to jointly advance the development of standardized real-machine benchmark frameworks, aiming to define the benchmark for real machine evaluation of embodied intelligence through deep collaboration among industry, academia, and research.



Upon its launch, RoboChallenge quickly garnered positive responses from the global embodied intelligence community. Open-source models such as $\pi 0$, $\pi 0.5$, RDT-1B,

CogACT, and OpenVLA-OFT, evaluated by the community and individuals, have successfully made the list. Currently, the teams from Spirit and Wall-oss have completed the full Table30 evaluation and entered the leaderboard. Meanwhile, more industry organizations are gradually joining the real machine evaluation: models from GigaBrain, BAAI, China Mobile (Hangzhou) Information Technology, Galaxea, and Horizon Robotics are also under ongoing evaluations. This trend of cross-border participation, gathering top research institutions, tech giants, unicorns, and open-source communities, represents a consensus in the embodied intelligence field to embrace real machine testing and standardized evaluation.

This report is based on deep analysis of tens of thousands of real machine testing data from RoboChallenge, presenting the current performance of VLA models in physical environments. However, this is merely the prologue to RoboChallenge's push for the standardization of real machine evaluation for embodied intelligence. In the future, we will continue to introduce more models, increase more diversified and realistic industrial scenario benchmark tasks, launch more challenging real machine evaluation tasks, and explore distributed real machine evaluation mechanisms. Our vision is to work with the community to create value for embodied intelligence in real physical environments through the most authentic real machine evaluation platform.

1. Core Findings and Highlights

- 1. The demand for testing has grown exponentially, making real machine validation a necessity in the industry.** Over the past three months, the number of registered users and evaluation submissions on the RoboChallenge platform has shown a significant exponential growth trend, indicating that RoboChallenge has become an important platform for evaluating the capabilities of embodied intelligence models.
- 2. Stacking bowls and moving objects into boxes have become "Hello World" level tasks.** Data shows that most submitting users have chosen these two tasks as their preferred testing tasks for VLA models, with a high success rate.
- 3. Organizing paper cups and making sandwiches remain challenging problems.** Tasks involving multi-step operations and fine manipulation are still too difficult for the current models, with low success rates, and some tasks even approaching 0.
- 4. The top model's success rate is about 50%, indicating room for improvement.** Currently, the highest success rate of models tested on the Table30 evaluation set is only about 50%, which still has room for enhancement and reflects the challenge of the Table30 task set.
- 5. VLA models are still working to master human intuitive operations.** Real testing data shows that while the tested models possess strong command semantic understanding abilities (showing a trend of improvement), their success rate in fine

manipulation tasks is below 15%. This phenomenon has accumulated a large amount of real machine failure data on the RoboChallenge platform, and this public 'failure data collection' can serve as a key reference for model iteration and optimization.

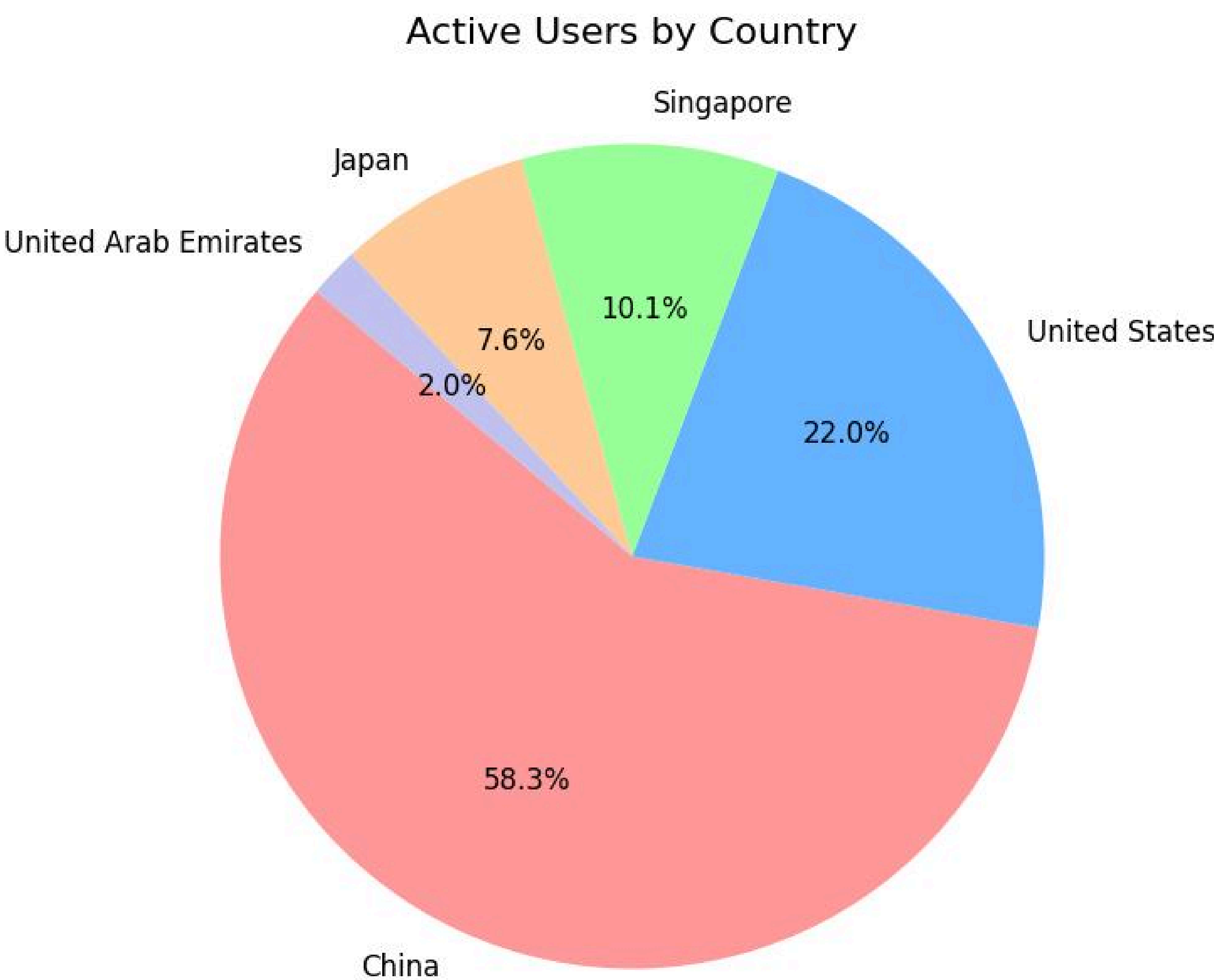
2. Data Insights

2.1 Platform User Data

Since the launch of the RoboChallenge platform, community participation has remained high, and the platform has accommodated a large-scale demand for evaluations during this period.

- **High testing conversion rate:** The platform has issued a total of 209 testing token applications. Among them, 82 developers have successfully run the local inference environment and submitted evaluation requests. The conversion rate from application to actual testing is 39.2%, reflecting the strong demand for standardized real machine evaluation in the field of embodied intelligence.
- **A large amount of evaluation data has been accumulated:** The total number of real machine tests (Rollouts) executed by the platform has exceeded 40,000 (41,969 times). The platform has formed a closed loop in the evaluation process, including task submission, real machine execution, log recording, and result publishing, providing a data foundation for further in-depth analysis.
- **High concurrent throughput capacity:** The platform achieved 181 model evaluation submissions in a single day (Runs), with a peak of 834 real machine tests (Rollouts) in a single day, validating the stability and reliability of the RoboChallenge platform architecture under high-load conditions.

2.2 Website Access Data Analysis



The distribution of active users by region shows that RoboChallenge is forming an international ecosystem. Although developers from China currently make up 58.3% of the core force, developers from the United States (22.0%), Singapore (10.1%), Japan (7.6%), and the UAE (2.0%) are also continuously joining, competing on the same standardized real machine evaluation platform.

■ 2.3 Open Source Platform Data Summary and Analysis

Since the launch of RoboChallenge, the platform has adhered to the principle of synchronous open-sourcing of evaluation and data, making datasets, task definitions, and supporting resources related to RoboChallenge available on the two major open-source platforms, GitHub and Hugging Face. This lowers the usage threshold for researchers and developers, promoting the reproducibility and comparability of real robot evaluations.

- Hugging Face: <https://huggingface.co/RoboChallenge>
- GitHub: <https://github.com/RoboChallenge>

On the Hugging Face platform (<https://huggingface.co/RoboChallenge>), RoboChallenge officially released the complete Table30 dataset, covering 30 standardized desktop operation tasks, supporting model training, offline evaluation, and result reproduction. As of now, the Table30 dataset has been downloaded a total of 17K times (7K times in the past month), indicating the community's ongoing interest and practical demand for a unified benchmark for real robot tasks. This dataset has been widely used for model capability comparisons, method ablation experiments, and research on sim-to-real transfer, gradually becoming one of the foundational reference resources in the field of embodied intelligence.

Meanwhile, on GitHub (<https://github.com/RoboChallenge>), the RoboChallenge open-source repository has concentrated on releasing evaluation codes, task configuration examples, and usage instructions related to the platform, facilitating researchers to quickly access the real robot evaluation process. It continues to receive issues and improvement suggestions from the community, reflecting that RoboChallenge is forming an early developer and researcher community centered around "real evaluation + open collaboration". Through dual-platform collaborative open sourcing on GitHub and Hugging Face, RoboChallenge not only achieves transparency in data, code, and evaluation processes but also lays the community foundation for the future integration of more models, tasks, and hardware platforms. This strategy allows RoboChallenge to enter the workflows of researchers and developers early, rather than remaining at the display level.

3. In-Depth Interpretation of the Leaderboard

As the world's first large-scale real machine evaluation platform for embodied intelligence, RoboChallenge's leaderboards have become an important reference for measuring the real machine performance of models. Since its release, the RoboChallenge leaderboards have quickly gained widespread influence in the open-source community and the industry due to their unified testing standards and transparent evaluation mechanisms. The ranking results have not only been shared by many developers but have also received recognition from several leading model teams, gradually evolving into an important evaluation platform for validating algorithm iteration effects and model performance.

Before delving into the analysis of the RoboChallenge leaderboards, it is essential to understand the entry thresholds and the transparent mechanisms involved. The RoboChallenge rankings are not a simple listing of scores but are based on the following three key core rules.

1. Overall Leaderboard Entry Threshold: Unlike leaderboard that test only a single skill (such as grasping), RoboChallenge adopts a more comprehensive evaluation standard. Only models that have fully completed the Table30 evaluation of all 30 tasks are eligible for the overall ranking. This means that every model on the list has undergone systematic testing from basic rigid body grasping to complex long-range and flexible object manipulation.
2. Transparent Evaluation Mechanism: RoboChallenge is committed to building a transparent evaluation ecosystem. Evaluation recordings and machine logs of participants are publicly released on the platform. Whether it is a perfect operation with a success rate of 100% or a failed attempt with a success rate of 0%, all data is open for global developers and users to review.
3. Systematic Elimination of Human Bias: To reduce the impact of operator subjectivity on evaluation results, RoboChallenge has introduced a visual input matching method. By overlaying a semi-transparent reference image on the operation side, it requires that the object placement be adjusted to closely match the reference scene in the training data before each test. This mechanism eliminates systematic biases caused by operators' subjective placements, such as unconsciously placing objects in more easily graspable "sweet spots", ensuring fair evaluations under as consistent initial conditions as possible.

■ 3.1 Overall Leaderboard

The RoboChallenge homepage overall leaderboard is sorted by success rate (Success Rate) from high to low, as we believe this metric best reflects the core capability of models to solve real-world problems.

To present model performance more precisely, RoboChallenge adopts a dual evaluation system of success rate and progress score, and has established strict quantitative calculation standards. To minimize the randomness of a single run, each model must perform 10 real machine tests (Rollouts) for each task in Table30.

- **Success Rate:**

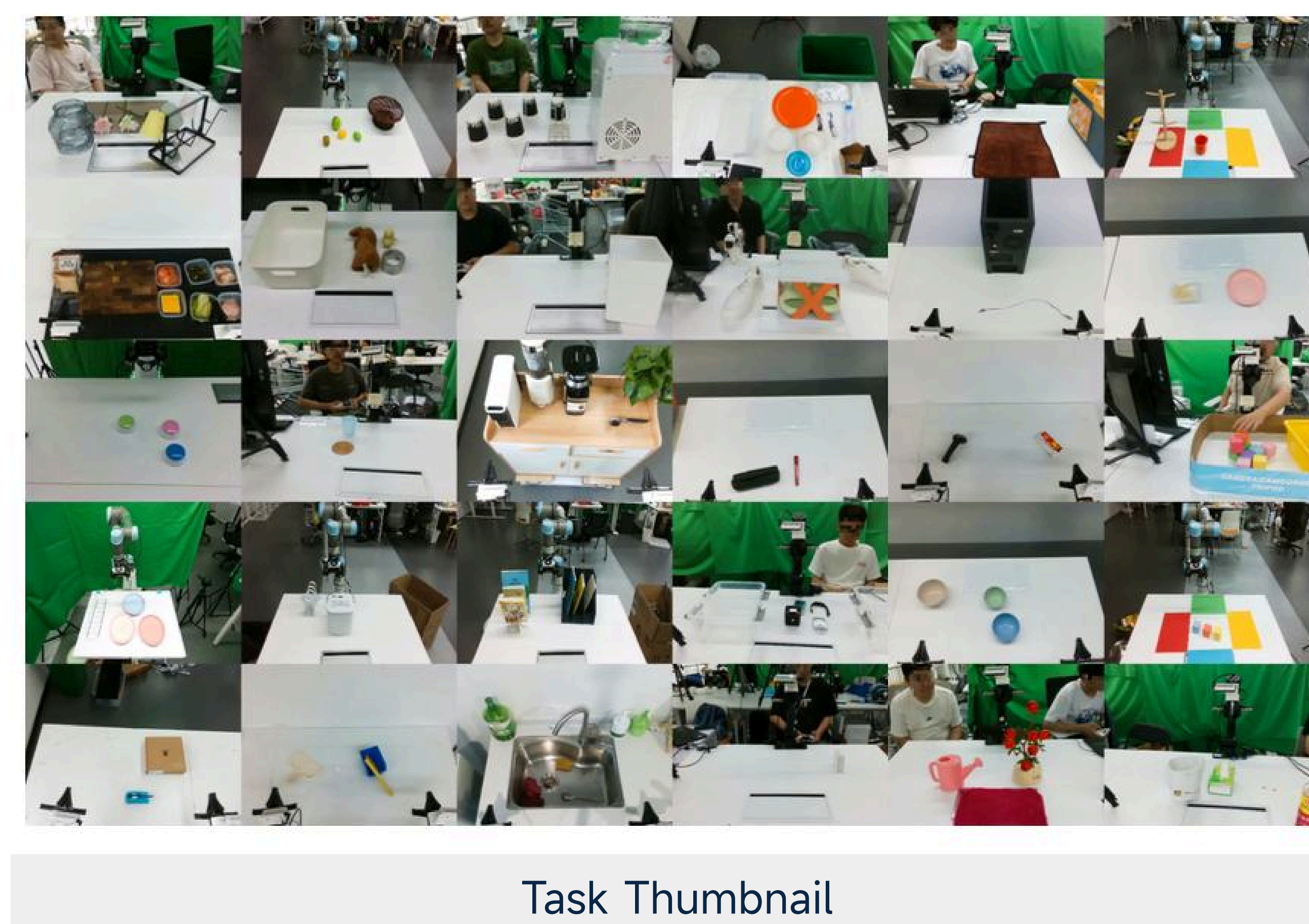
- Single Task Calculation: The proportion of successful attempts in 10 executions. For example, if successful 5 times, the success rate for that task is 50%.
- Overall Leaderboard Calculation: The average success rate across all 30 tasks.

- **Progress Score:**

- Single Task Calculation: Each execution has a full score of 10 points (0.5 points deducted for each retry), and the scores from 10 executions are accumulated to determine the final score for that task (maximum score of 100 points).
- Overall Leaderboard Calculation: The average of the task scores (each 0–100 points) across all 30 tasks.

Rank	Model/User	Score	Success Rate
1	Spirit-v1.5/Spirit AI 🏆	67.19	51.00%
2	pi0.5/rc_baseline	61.84	42.67%
3	wall-oss-v0.1/Pushi Zh...	55.30	35.33%
4	pi0/rc_baseline	46.41	28.33%
5	pi05_generalist/wyf	31.27	17.67%
6	RDT-10/zsz	28.84	15.00%
7	cogact/hsk	21.83	11.67%
8	pi0_generalist/wyf	20.22	9.00%

Homepage Leaderboard Ranking (Top 8 only)
Screenshot Date: 2026.1.23



Data observation

- Challenge of Table30 Testing: The average success rate of the top model is 51%. This data objectively reveals the rigor of real machine evaluations; even the best-performing model in this assessment maintains an end-to-end execution success rate of only about 50% when facing the comprehensive tasks of rigid bodies, soft bodies, and long-range covered by Table30.
- Process reveals potential capabilities: Data analysis shows that the scores of various models (Score) are generally higher than their success rates (e.g., Rank 2 model: Score 61.84 > SR 42.67%). The tested models are not entirely "ineffective" in failed tasks; rather, in many cases, they have completed more than half of the critical steps.
- There is room for improvement in model capabilities: The current highest success rate of 51% indicates that embodied intelligent models still have significant potential for improvement in real machine evaluation tasks. We believe that there are still many models in the industry that have not yet participated in evaluations. RoboChallenge sincerely invites global developers to participate in evaluations, showcasing stronger model capabilities on this fair, standardized platform, and continuously refreshing the highest success rate together.

3.2 Single/Multi-task Model Rankings

The RoboChallenge platform distinguishes between single-task models and multi-task models in the evaluation submission process, explained as follows:

- Single-task Model (Task-Specific): A model fine-tuned or trained for a single specific task.
- Multi-task Model (Multitask): A model that can handle 2-30 different tasks in Table30.

model type *

Multitask

Task-Specific

selected tasks *

Table 30

arrange_flowers

clean_dining_table

make_vegetarian_sandwich

place_shoes_on_rack

arrange_fruits_in_basket

fold_dishcloth

move_objects_into_box

plug_in_network_cable

arrange_paper_cups

hang_toothbrush_cup

open_the_drawer

pour_fries_into_plate

To facilitate targeted analysis for users and researchers, the Leaderboard page on the RoboChallenge official website provides filtering options, allowing users to view the rankings of single-task models and multi-task models separately.

Rank	Model/User	Is multitask	Score	Success Rate
1	Spirit-v1.5/Spirit AI	x	67.19	51.00%
2	pi0.5/rc_baseline	x	61.84	42.67%
3	wall-oss-v0.1/Pushi Zhang	x	55.30	35.33%
4	pi0/rc_baseline	x	46.41	28.33%
5	RDT-1B/zsz	x	28.84	15.00%
6	cogact/hsk	x	21.83	11.67%
7	openvla-oft/gkf	x	8.66	5.00%

Single-task Leaderboard

Rank	Model/User	Is multitask	Score	Success Rate
1	pi05_generalist/wyf	✓	31.27	17.67%
2	pi0_generalist/wyf	✓	20.22	9.00%

Multi-task Leaderboard

Comparison of the same base model's performance under single-task and multi-task settings:

- pi0.5: Dropped from 42.67% in the single-task model to 17.67% in the multi-task model (pi05_generalist), a decrease of about 25 percentage points.
- pi0: Dropped from 28.33% in the single-task model to 9.00% in the multi-task model (pi0_generalist), a decrease of nearly 20 percentage points.

Developer Invitation: Challenge the Baseline, Define New Heights

It should be noted that the only two models currently on the multi-task leaderboard (i.e., pi05_generalist and pi0_generalist) are benchmark models trained by the RoboChallenge team. These models were strictly limited in data scale during the training phase (approximately 50 samples per task), with the main goal of providing the community with a reference score for benchmarking.

Therefore, the current multi-task leaderboard results more accurately represent the lower performance bound under small-sample multi-task mixed training scenarios, rather than the maximum performance that multi-task embodied intelligence models can potentially reach. We sincerely invite global developers to explore and improve upon this baseline, together advancing the capabilities of multi-task embodied intelligent models to new heights.

■ **3.3 Leaderboard Across Different Dimensions**

To break through the limitations of a single score metric and evaluate model capabilities in a more dimensional and granular way, RoboChallenge has constructed a multi-dimensional task labeling system around the Table30 evaluation set.

3.3.1 Mapping of Tasks and Labels

Each task in Table 30 is not isolated; it is assigned three-dimensional attribute labels: model type, configuration (Arm), and capability type.

Dimension	Tag	Description	Task Count	Typical Task Example
Robot Type	ALOHA	–	11	Pour fries (pour_fries_into_plate)
	ARX5	–	11	Place shoes on rack (place_shoes_on_rack)
	UR5	–	6	Shred paper (shred_scrap_paper)
	Franka	–	2	Press buttons (press_three_buttons)
Configuration	Single-arm	Single-arm operation	19	Put cup on coaster (put_cup_on_coaster)
	Two-arm	Dual-arm operation	11	Stick tape (stick_tape_to_box)
Task Type	Precise3d	High-precision 3D operation	12	Plug in network cable (plug_in_network_cable)
	Repeated	Repetitive task	10	Arrange flowers (arrange_flowers)
	Bimanual	Dual-arm/hand coordination	8	Sweep rubbish (sweep_the_rubbish)
	Manipulation	General manipulation/grasping	6	Open drawer (open_the_drawer)
	Classification	Visual classification	5	Sort electronics (sort_electronic_products)
	Multiview	Multi-view perception	5	Search green boxes (search_green_boxes)
	Simple-pick	Simple grasping	4	Place shoes on rack (place_shoes_on_rack)
	Temporal	Sequential/Long-horizon memory	3	Scan QR code (scan_QR_code)
	Softbody	Deformable/Soft object	3	Fold dishcloth (fold_dishcloth)

3.3.2 Performance by Task

Performance of the Top 9 models on different tasks:

Task	Spirit-v1.5		pi0.5		wall-oss-v0.1		pi0		pi05_generalist		RDT-1B		cogact		pi0_generalist		openvla-oft	
	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score
Average	51	67.19	42.67	61.84	35.33	55.3	28.33	46.41	17.67	31.27	15	28.84	11.67	21.83	9	20.22	5	8.66
arrange flowers	50	78.5	50	69.5	20	67	50	67.5	0	30.5	10	43	10	22.5	0	13.5	0	0
arrange fruits in basket	80	93.5	40	70.5	80	82	20	22.5	0	9	0	5	80	88	0	11.5	0	0
arrange paper cups	0	56	0	48	0	48.5	0	41.5	0	31	10	50	0	8.5	0	15	0	5
clean dining table	30	51.5	10	58.5	10	38	0	33.5	30	62	0	3.5	0	4.5	0	25.5	0	0
fold dishcloth	20	20	20	24	10	41	0	32	0	0	30	34	0	0	0	0	0	0
hang toothbrush cup	80	88	50	71	60	74	50	70	50	71	0	0	30	65	20	62	0	19
make vegetarian sandwich	0	27.5	0	29.5	0	2	0	17.5	0	0	0	15.5	0	0	0	0	0	0
move objects into box	80	88.5	50	63.5	60	70.5	50	66	20	40	50	73.5	60	64.5	20	44.5	0	0
open the drawer	70	73.5	40	60.5	70	84.5	0	50	50	80	70	86	0	50	0	20	30	64
place shoes on rack	90	85	90	90.5	60	75.5	80	77	0	20	60	73	0	5	0	16.5	0	9
plug in network cable	0	24.5	20	65	0	10	20	45	0	0	0	4	0	6.5	0	0	0	0
pour fries into plate	50	61	30	38	10	37	40	56	0	0	10	34	0	23	0	0	0	0
press three buttons	90	96	0	0	100	100	0	0	0	4	0	0	0	18	0	0	0	0
put cup on coaster	90	96	90	96	70	88	60	71	70	63	80	92	20	18	0	0	50	60
put opener in drawer	80	79.5	80	77.5	70	91	50	71.5	20	38	20	43	0	12	0	0	0	0
put pen into pencil case	90	85	80	89.5	70	81	70	88	50	63.5	0	28	20	30	0	14.5	0	0
scan QR code	0	61	50	55	20	52	30	30.5	0	7	0	0	0	4	0	3	0	0
search green boxes	90	87.5	80	80	50	49.5	70	74	0	3	10	28.5	30	33.5	0	0	0	0
set the plates	80	77	80	88	50	65	10	34.5	40	49.5	0	0	0	0	50	69.5	0	0
shred scrap paper	20	39	0	36	0	30.5	30	59	20	36	0	4	10	43	20	27	0	0
sort books	0	12	0	60	0	14	0	24.5	0	24	0	4.5	0	9.5	10	26.5	0	0
sort electronic products	30	46.1	50	68.6	0	23.1	0	31.1	0	22.5	0	18.6	0	0	0	22.5	0	3.4
stack bowls	100	98	100	99.5	70	76	100	98.5	80	83	50	55	10	13.5	40	53.5	0	2
stack color blocks	80	85	100	99	100	100	70	72.2	10	30	10	35	40	36	30	39	0	0
stick tape to box	20	38	10	29	10	52	10	28	0	16	0	10	0	0	0	0	0	0
sweep the rubbish	60	74.5	20	46	10	33.5	10	27	10	46	0	0	0	8.5	0	17	0	0
turn on faucet	70	72	100	99	20	44	20	23	60	56	20	45	10	34.5	60	67.5	60	68.5
turn on light switch	80	85	40	61	40	63	10	40	10	25	20	32	30	48	20	29	10	29
water potted plant	0	85	0	36.5	0	41	0	6	0	0	0	11	0	9	0	0	0	0
wipe the table	0	51.5	0	46	0	25.5	0	35	10	28	0	37	0	0	0	29	0	0

Data Observation

- **Spirit-v1.5:** Achieved the highest success rate and highest process score (including ties, SR > 0) across 9 tasks.
 - Highlight task: Moving objects into box (move_objects_into_box) with a success rate of SR 80% / Score 88.5.
 - Task labels: repeated, single-arm, Franka
- **pi0.5:** Achieved the highest success rate and highest process score (including ties, SR > 0) across 7 tasks.
 - Highlight task: Stacking bowls (stack_bowls) with a success rate of SR 100% / Score 99.5, this task is also the most frequently tested on the current platform.
 - Task labels: bimanual, two-arm, repeated, classification, ALOHA
- **wall-oss-v0.1:** Achieved a perfect score of SR 100% / Score 100 on 2 specific tasks.
 - Highlight tasks: Pressing three buttons (press_three_buttons) and stacking color blocks (stack_color_blocks).
 - Task labels:
 - temporal, single-arm, Franka, repeated
 - simple-pick, single-arm, UR5, classification

3.3.3 Performance by Robot Type

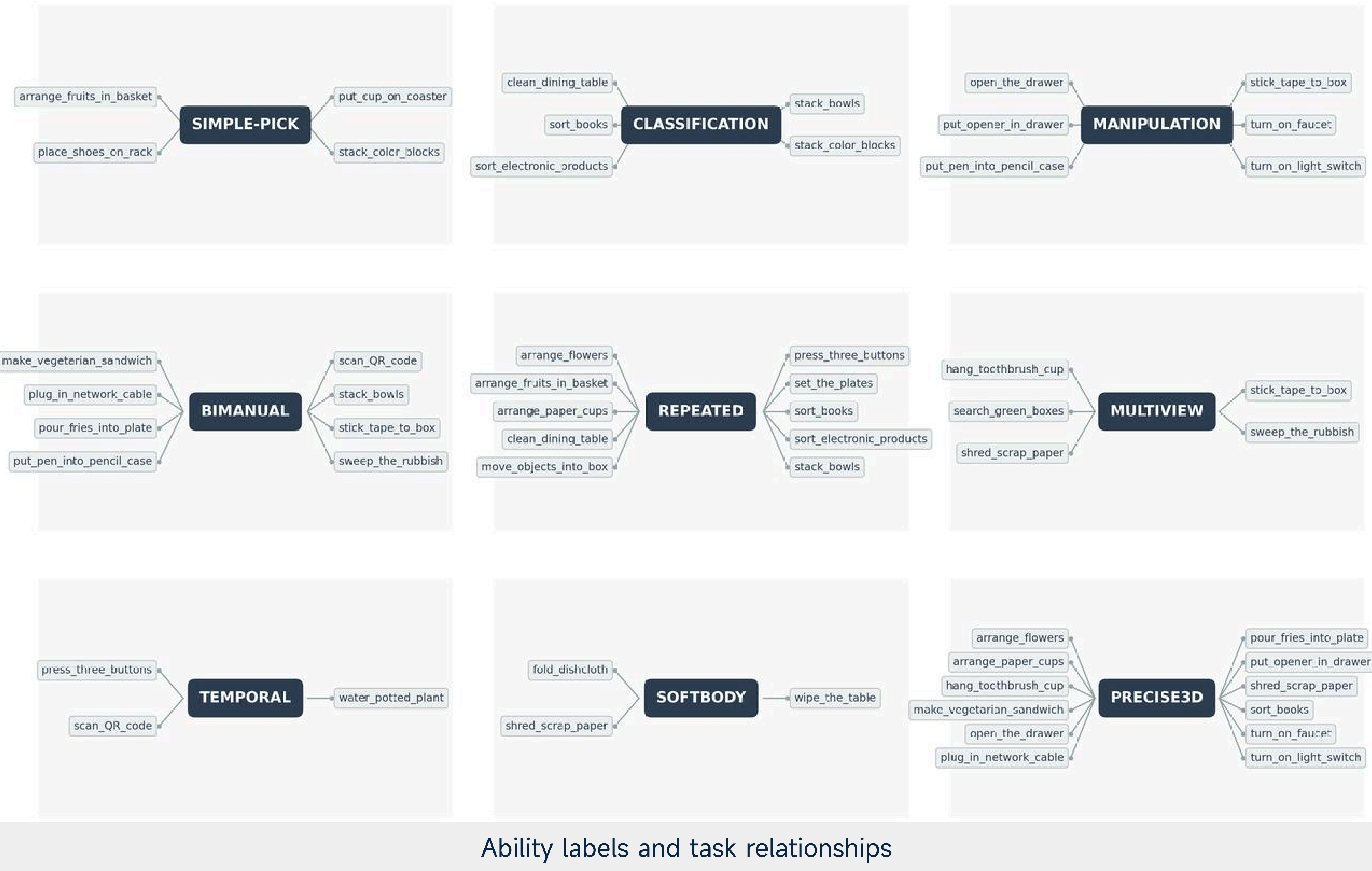
Table 30 evaluation set includes four mainstream robot hardware types: ALOHA, ARX5, UR5, and Franka, covering both single-arm and dual-arm configurations. Performance of the Top 9 models across different robot types:

Tag	Tasks	Spirit-v1.5	pi0.5	wall-oss-v0.1	pi0	pi05_generalist	RDT-1B	cogact	pi0_generalist	openvla-oft
ALOHA	11	45.5	45.5	26.4	31.8	22.7	9.1	3.6	9.1	5.5
ARX5	11	47.3	41.8	29.1	24.5	12.7	26.4	8.2	1.8	8.2
Franka	2	85	25	80	25	10	25	30	10	0
UR5	6	56.7	45	48.3	30	20	1.7	26.7	21.7	0

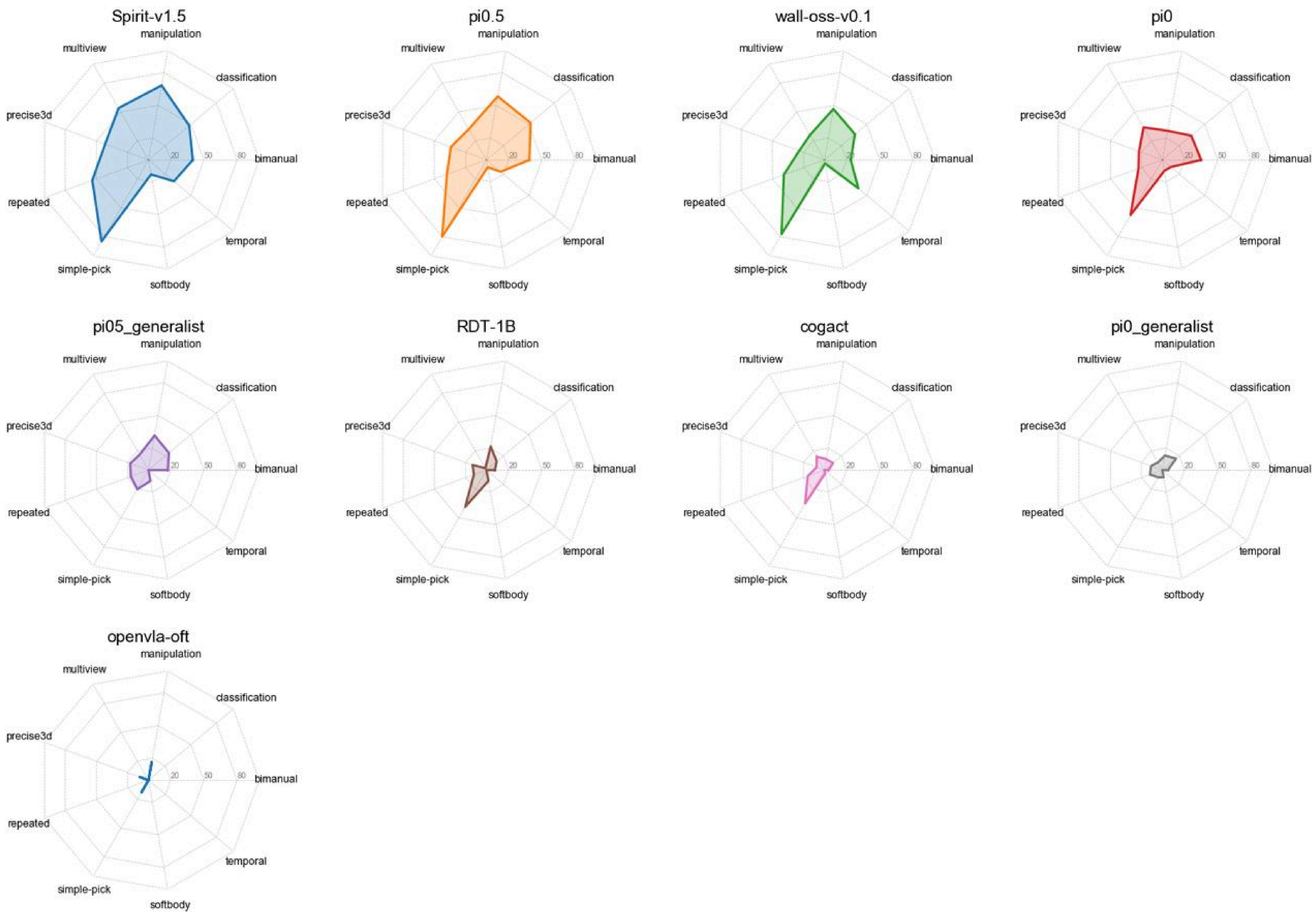
3.3.4 Performance by Capability Label

Through analysis based on task capability labels, we can see the specific performance (success rates) of the models that made the list across different capability dimensions.

Tag	Tasks	Spirit-v1.5	pi0.5	wall-oss-v0.1	pi0	pi05_generalist	RDT-1B	cogact	pi0_generalist	openvla-oft
bimanual	8	40	38.8	23.8	35	17.5	7.5	3.8	5	0
classification	5	48	52	36	34	24	12	10	16	0
manipulation	6	68.3	58.3	46.7	26.7	31.7	21.7	10	13.3	16.7
multiview	5	54	32	26	34	16	2	14	8	0
precise3d	12	41.7	34.2	24.2	22.5	17.5	13.3	7.5	10.8	8.3
repeated	10	54	38	39	23	17	12.1	16	12	0
simple-pick	4	85	80	77.5	57.5	20	38.5	35	7.5	12.5
softbody	3	13.3	6.7	3.3	10	10	10	3.3	6.7	0
temporal	3	30	16.7	40	10	0	0	0	0	0



Model Capabilities (SR%)



Data observation

- **Simple-pick** (such as tasks like stack_color_blocks, put_cup_on_coaster): This is currently the best-performing area for the listed models, with Spirit-v1.5 (85%), pi0.5 (80%), and wall-oss-v0.1 (77.5%) all significantly exceeding 50%.
- **Manipulation** (such as tasks like stick_tape_to_box, open_the_drawer): Spirit-v1.5 achieved a success rate of 68.3% in these tasks, leading among the tested models.
- **Classification** (such as tasks like sort_books, stack_color_blocks): pi0.5 surpassed Spirit-v1.5 (48%) with a success rate of 52%.
- **Softbody** (such as tasks like wipe_the_table, fold_dishcloth): A challenging area in the Table30 tasks. Even the best-performing Spirit-v1.5 only achieved a success rate of 13.3%, with other models generally below 10%.
- **Temporal** (such as tasks like press_three_buttons, scan_QR_code): Models for tasks involving long-term memory generally performed poorly, with wall-oss-v0.1 achieving a success rate of 40% in these tasks.

4. Task Difficulty Analysis

The RoboChallenge leaderboard includes only models that have completed evaluations on all 30 tasks in the Table30 benchmark. To better highlight task-level performance, we additionally select the top three models for each individual task based on their best achieved scores.

4.1 Task Success Rate Analysis

By analyzing the success rate distribution of the top 3 models, we categorized these 30 standardized tasks into three different ability tiers.

Task Name	Rank 1 Model	Rank 1 SR(%)	Rank 1 Score	Rank 2 Model	Rank 2 SR(%)	Rank 2 Score	Rank 3 Model	Rank 3 SR(%)	Rank 3 Score
arrange flowers	Spirit-v1.5	50	78.5	pi0.5	50	69.5	pi0	50	67.5
arrange_fruits_in_basket	Spirit-v1.5	80	93.5	cogact	80	88	wall-oss-v0.1	80	82
arrange paper cups	rdtl	20	48.5	RDT-1B	10	50	Spirit-v1.5	0	56
clean dining table	pi05_generalist	30	62	Spirit-v1.5	30	51.5	pi0.5	10	58.5
fold dishcloth	RDT-1B	30	34	model_test	20	27	pi0.5	20	24
hang_toothbrush_cup	Spirit-v1.5	80	88	Model_Test	70	82	wall-oss-v0.1	60	74
make vegetarian sandwich	GigaBrain-0.1	0	31.5	pi0.5	0	29.5	Spirit-v1.5	0	27.5
move_objects_into_box	test	90	88.5	GigaBrain-0.1	90	86.5	Spirit-v1.5	80	88.5
open_the_drawer	RDT-1B	70	86	wall-oss-v0.1	70	84.5	Spirit-v1.5	70	73.5
place shoes on rack	pi0.5	90	90.5	Spirit-v1.5	90	85	pi0	80	77
plug_in_network_cable	pi0.5	20	65	pi0	20	45	GigaBrain-0.1	0	50
pour fries into plate	GigaBrain-0.1	50	74	Spirit-v1.5	50	61	VA_Test	40	66
press_three_buttons	wall-oss-v0.1	100	100	Spirit-v1.5	90	96	wx-test	70	79
put_cup_on_coaster	pi0.5	90	96	Spirit-v1.5	90	96	GigaBrain-0.1	90	88.5
put_opener_in_drawer	Spirit-v1.5	80	79.5	pi0.5	80	77.5	wall-oss-v0.1	70	91
put_pen_into_pencil_case	Spirit-v1.5	90	85	pi0.5	80	89.5	pi0	70	88
scan_QR_code	pi0.5	50	55	pi0	30	30.5	wall-oss-v0.1	20	52
search green boxes	Spirit-v1.5	90	87.5	pi0.5	80	80	pi0	70	74
set the plates	GigaBrain-0.1	90	88.5	pi0.5	80	88	Spirit-v1.5	80	77
shred scrap paper	pi0	30	59	Spirit-v1.5	20	39	pi05_generalist	20	36
sort books	pi0_generalist	10	26.5	pi0.5	0	60	GigaBrain-0.1	0	27
sort electronic products	pi0.5	50	68.6	Spirit-v1.5	30	46.1	omnivla-0	0	35.6
stack bowls	pi0.5	100	99.5	pi0	100	98.5	Model_Test	100	98.5
stack color blocks	wall-oss-v0.1	100	100	GigaBrain-0.1	100	100	pi0.5	100	99
stick_tape_to_box	GigaBrain-0.1	60	71	Model_VLA	30	62	Model_Test	30	46
sweep_the_rubbish	Spirit-v1.5	60	74.5	VLA_Test	60	69	pi0.5	20	46
turn_on_faucet	VLA-Eval	100	100	pi0.5	100	99	Spirit-v1.5	70	72
turn_on_light_switch	Spirit-v1.5	80	85	wall-oss-v0.1	40	63	model-test-0108	40	62
water potted plant	GigaBrain-0.1	60	86.5	Spirit-v1.5	0	85	wall-oss-v0.1	0	41
wipe_the_table	pi05_generalist	10	28	Spirit-v1.5	0	51.5	pi0.5	0	46

4.1.1 Tier 1 : Hello World Level Tasks

Definition: For top-performing models, these tasks are no longer a bottleneck: the top three models reach a 100% success rate, indicating that the underlying skills have been effectively mastered at the current frontier.

Representative Task:

- Stacking bowls (stack_bowls): Top 1/2/3 success rates are all 100%, making it the most frequently tested "Hello World" level task.
- Stack color blocks (stack_color_blocks): Top 1/2/3 success rates are all 100%.

4.1.2 Tier 2 : Simple Tasks

Definition: These tasks are relatively friendly to most top models. The top 1 success rate exceeds 90%, and the lowest success rate among the top 3 is no less than 70%, indicating that these tasks have a low threshold and are easy to complete.

Representative Tasks:

- Placing shoes on a rack (place_shoes_on_rack): Top 1 (pi0.5) success rate 90%, top 3 is 80%.
- Searching for green boxes (search_green_boxes): Top 1 (Spirit-v1.5) success rate 90%, top 3 is 70%.

4.1.3 Tier 3: Specialties of Specific Models

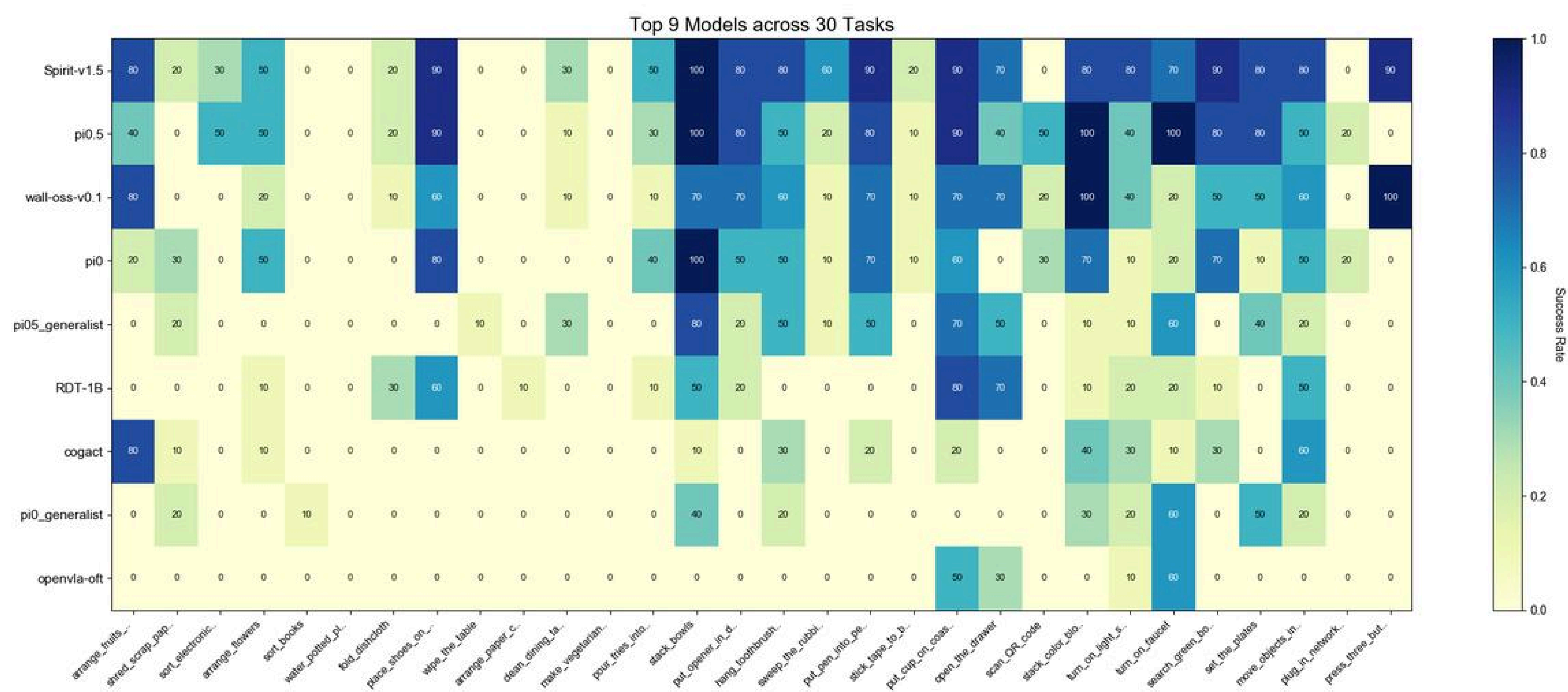
Definition: These tasks exhibit extreme polarization. The top 1 model performs perfectly (100% success rate), but the subsequent models show a significant drop in performance.

Representative Tasks:

- Pressing three buttons (press_three_buttons):
 - Top 1 (wall-oss-v0.1) success rate: 100%
 - Top 3 success rate: drops to 70%

4.2 Failure Tasks Analysis

In the heatmap of evaluation results for the listed models (task dimension), there are some tasks that not only have a low average success rate but also repeatedly show a 0% success rate for all tested models. For example, tasks like making a vegetarian sandwich (make_vegetarian_sandwich) and watering a potted plant (water_potted_plant) are vividly referred to as the "Wall of Sighs".



Below is an analysis of five failure tasks:

4.2.1 Failure Case One: make_vegetarian_sandwich

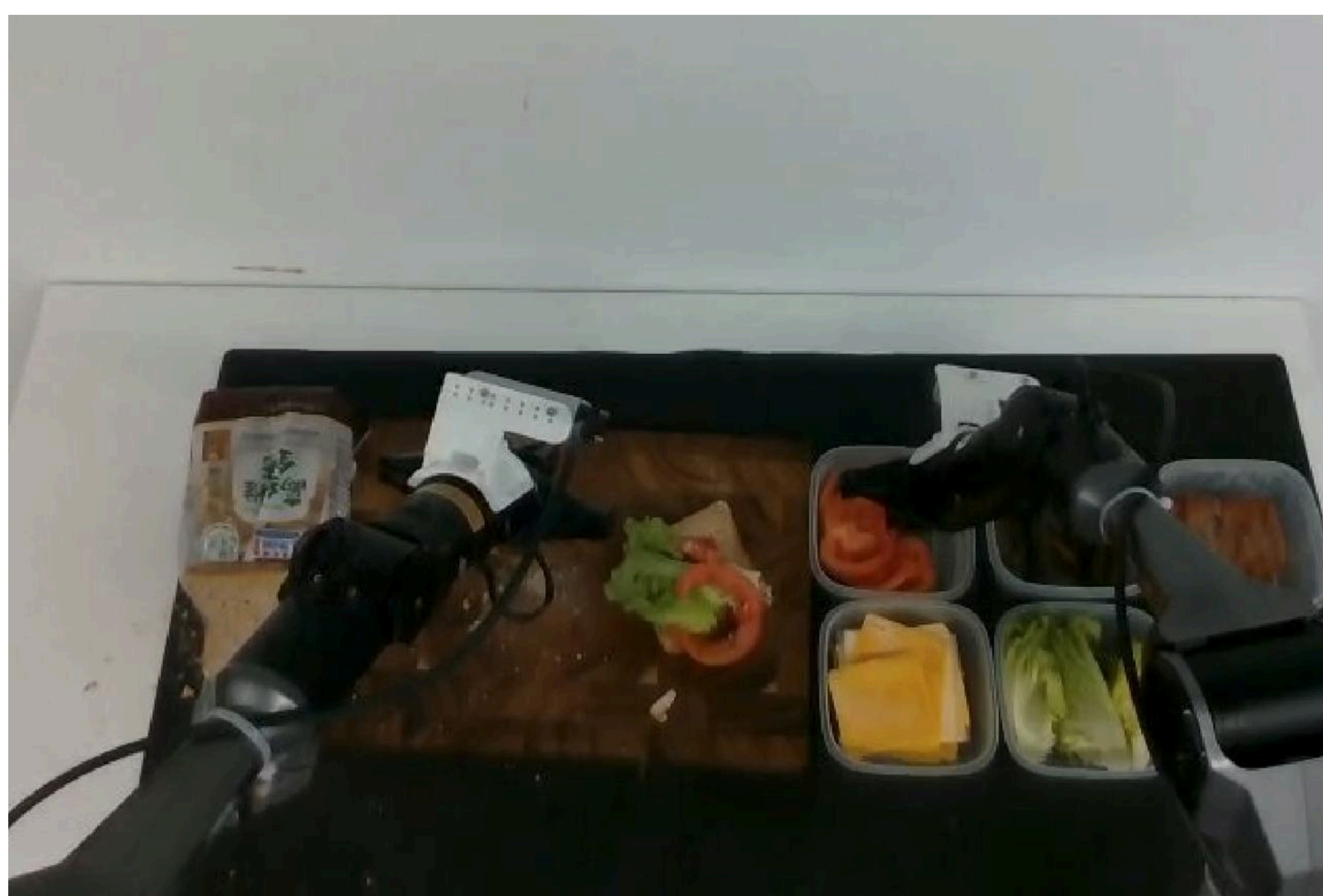
Tags: bimanual, two-arm, ALOHA, precise3d

Scoring Points:

- 2.0 points: Place the bread.
- 2.0 points: Place the vegetables.
- 2.0*2 points: Place the tomatoes.
- 2.0 points: Place the bread.

Data: All listed models have a success rate of 0%.

Screenshot:



Review: This task requires a strict sequence (place bread -> vegetables -> tomatoes -> bread). Evaluations show that models often make errors or fail at the first step (grasping the item with the left arm).

Attribution: "One step wrong, every step wrong". This type of task has an extremely low tolerance for error; if the initial step (like not grasping the item or picking the wrong one) fails, the entire task is declared a failure.

4.2.2 Failure Case Two: water_potted_plant

Tags: temporal, single-arm, ARX5

Scoring Points:

- 4.0 points: Precisely grasp the handle of the kettle.
- 2.0 points: Move the kettle to the position of the potted plant.
- 3.0 points: Water the potted plant using the kettle.
- 1.0 points: Place the kettle back to its original position.

Data: The success rate of all listed models is 0%.

Screenshot:



Review: The model successfully completed the initial actions (grabbing the kettle, moving to the potted plant), but after repeatedly attempting to water for a while, the robotic arm suddenly exhibited abnormal behavior—extending and moving significantly, and failed to execute the termination action of placing the kettle back in its original position.

Attribution: Lack of temporal dependency. Long-term tasks require the model to maintain memory of historical states. Once there is a loss of state in the intermediate stages, the model can become logically confused, resulting in random actions akin to "hallucinations".

4.2.3 Failure Case Three: sort_books

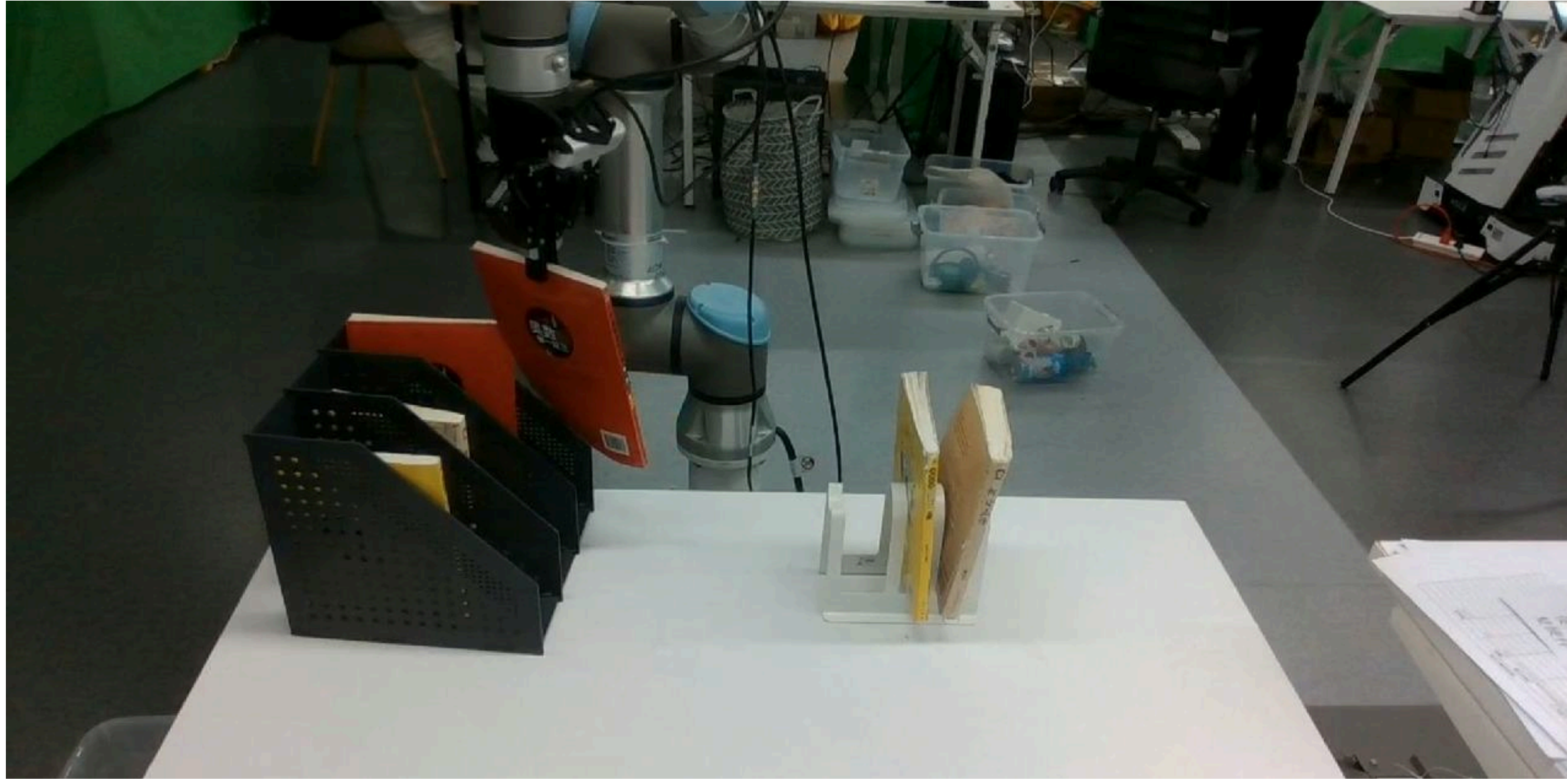
Tags: precise3d, single-arm, UR5, repeated, classification

Scoring Points:

- 1.5 * 3 points: Pick up one book.
- 1.5 * 3 points: Place one book in the corresponding position.
- 1.0 points: Reset the robotic arm.

Data: The highest success rate of the listed models is only 10%.

Screenshot:



Review: After failing to grasp the first book, the model skipped directly to attempting to grasp the second book. However, during execution, the robotic arm accidentally grasped the third book, leading to complete chaos in the task.

Attribution: Insufficient visual perception accuracy and error accumulation. Such tasks require extremely high visual segmentation capability (distinguishing closely placed books). Once there is a slight deviation in visual judgment, or if the previous operation fails, the error can be magnified over multiple operations, ultimately lowering the overall success rate.

4.2.4 Failure Case Four: fold_dishcloth

Tags: softbody, single-arm, ARX5

Scoring Points:

- 4.0 points: Fold the dishcloth the first time.
- 4.0 points: Fold the dishcloth the second time.
- 2.0 points: Place the folded dishcloth in the front-left position.

Data: The highest success rate of the listed models is 30%.

Screenshot:



Review: When the robotic arm attempted to fold the dishcloth, the fabric underwent unpredictable changes in shape due to stress, causing multiple attempts to fail to complete the fold, with the most common outcome being dragging the dishcloth along.

Attribution: Difficulty in predicting physical deformation. Compared to rigid bodies, the difficulty of operating soft bodies increases exponentially. Current algorithms struggle to accurately control the force and grasping position for flexible objects, which is a well-recognized challenge in the industry.

4.2.5 Failure Case Five: `arrange_paper_cups`

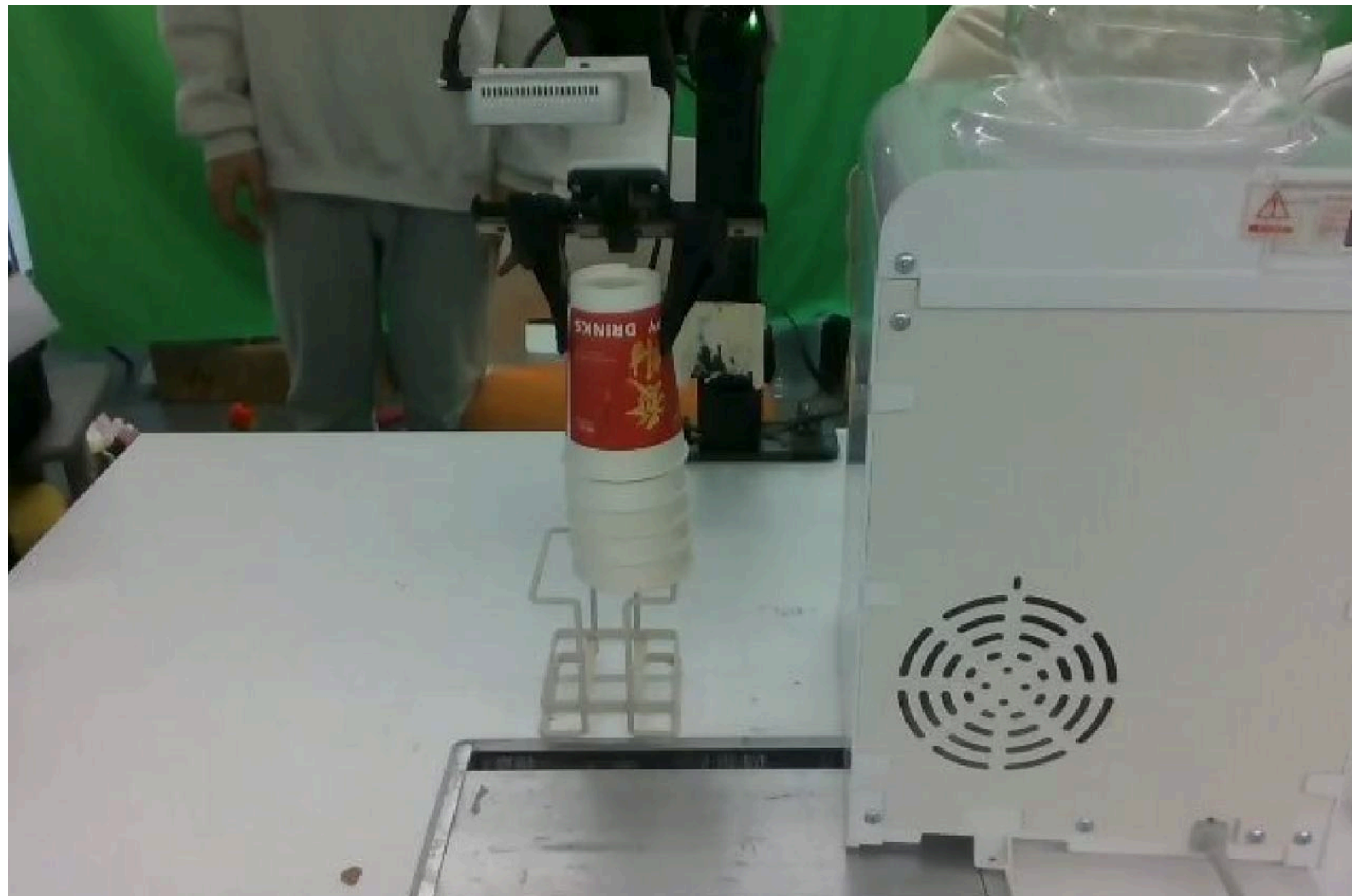
Tags: repeated, single-arm, ARX5, precise3d

Scoring Points:

- 1.0 * 4 points: Pick up each cup.
- 1.0 * 4 points: Place a cover on each cup.
- 2.0 points: Place the stacked cups on the shelf.

Data: The best model in the field has a success rate of 20%.

Screenshot:



Review: The model had actually successfully completed the first 80% of the work (successfully grasping and stacking 4 paper cups). However, in the final step—attempting to place the stacked cups on the shelf—the robotic claw accidentally knocked over the cup tower. Subsequently, the robotic arm made multiple attempts to re-grasp the fallen cups on the table, but all failed.

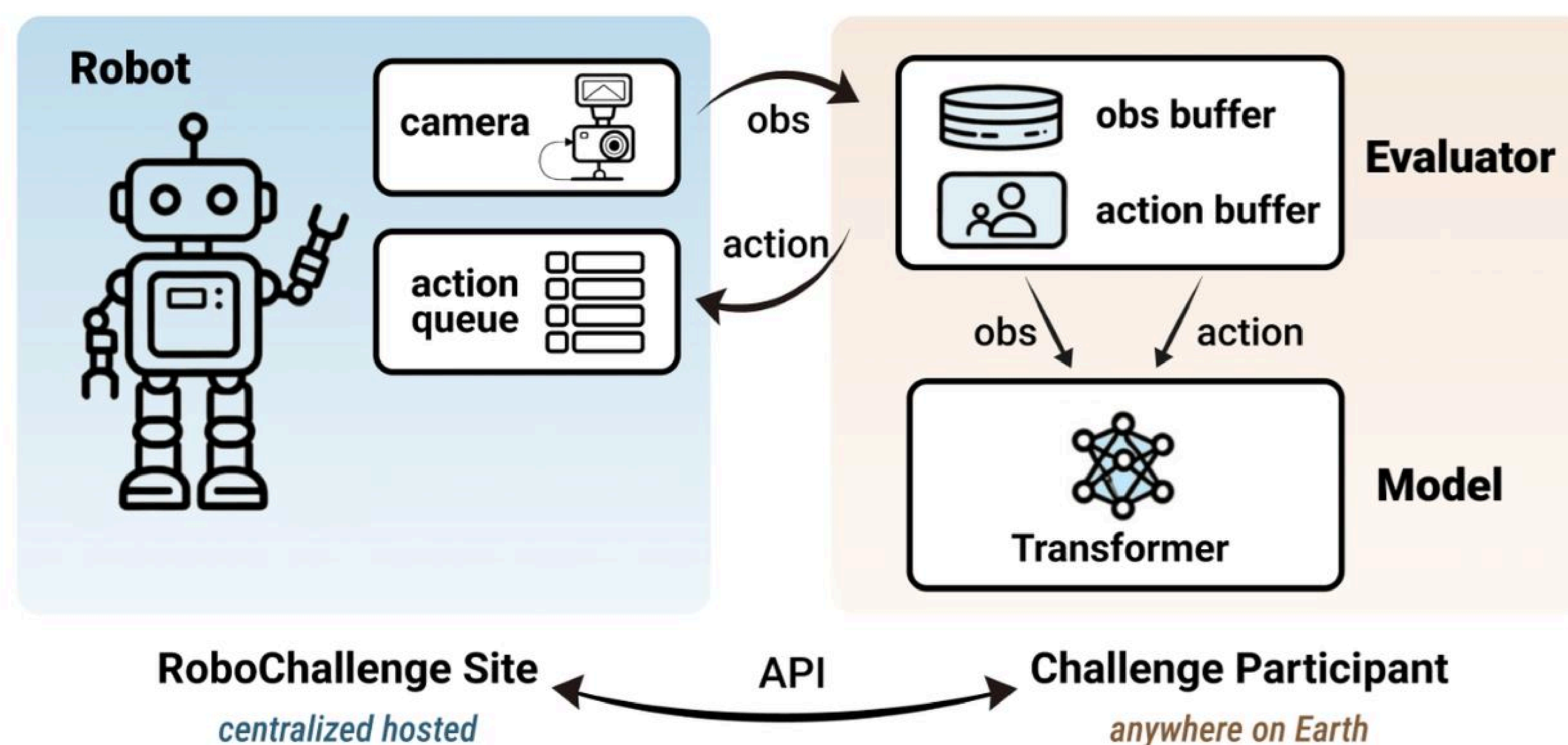
Attribution: Complex stacking tasks require perfect control throughout the process; any small shake in an action (like releasing the claw) can destroy all previous efforts.

5. RoboChallenge Evaluation System and Standardization Process

In the field of embodied intelligence, the standardization and fairness of real machine evaluations have long been a key bottleneck restricting horizontal comparisons of technology. RoboChallenge, the world's first large-scale real machine evaluation platform for embodied intelligence jointly initiated by Dexmal and Hugging Face, has established a simple and user-friendly remote real machine evaluation system. Next, we will detail its architectural design and scoring criteria.

■ 5.1 API-Based Remote Real Machine Evaluation Platform

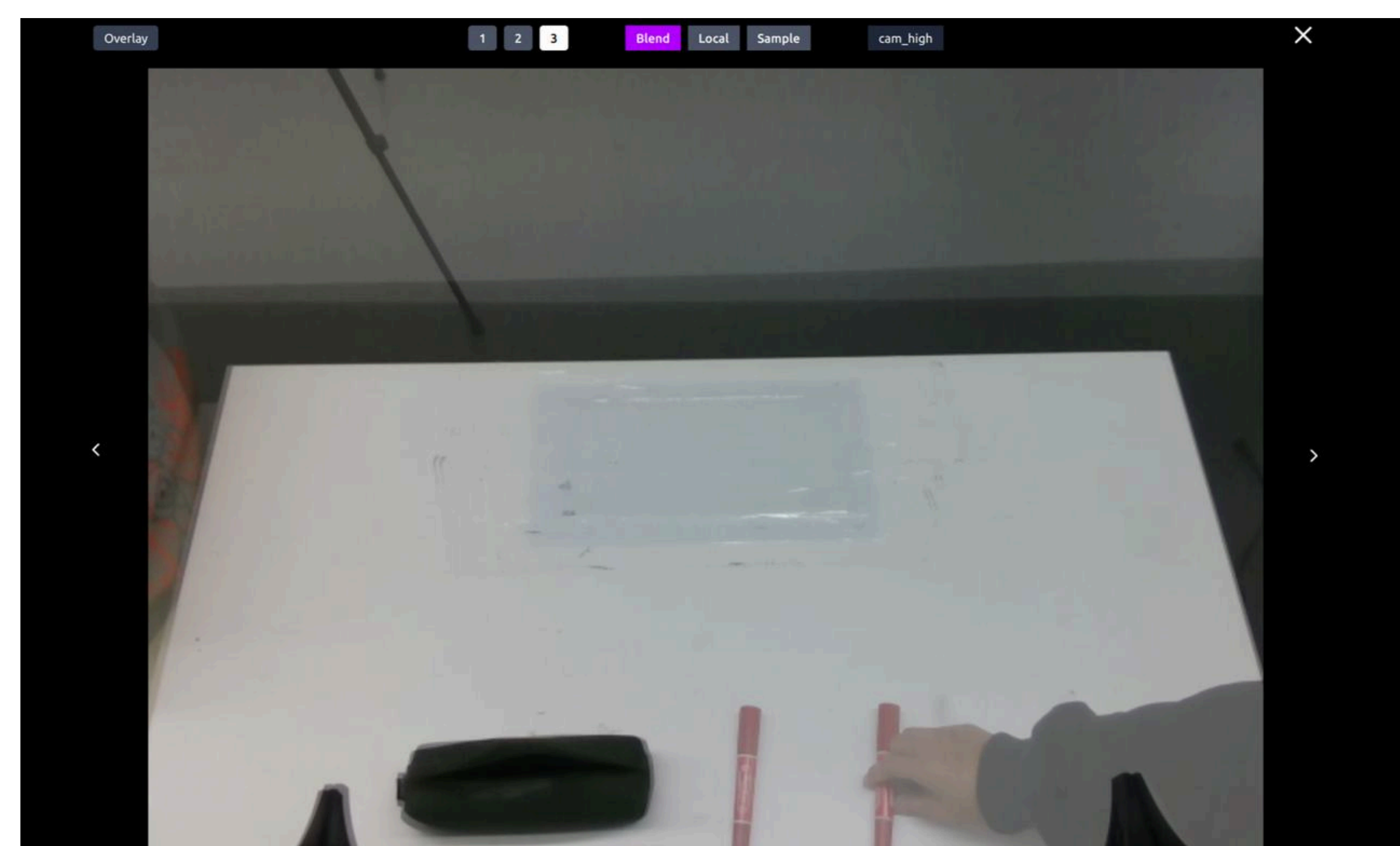
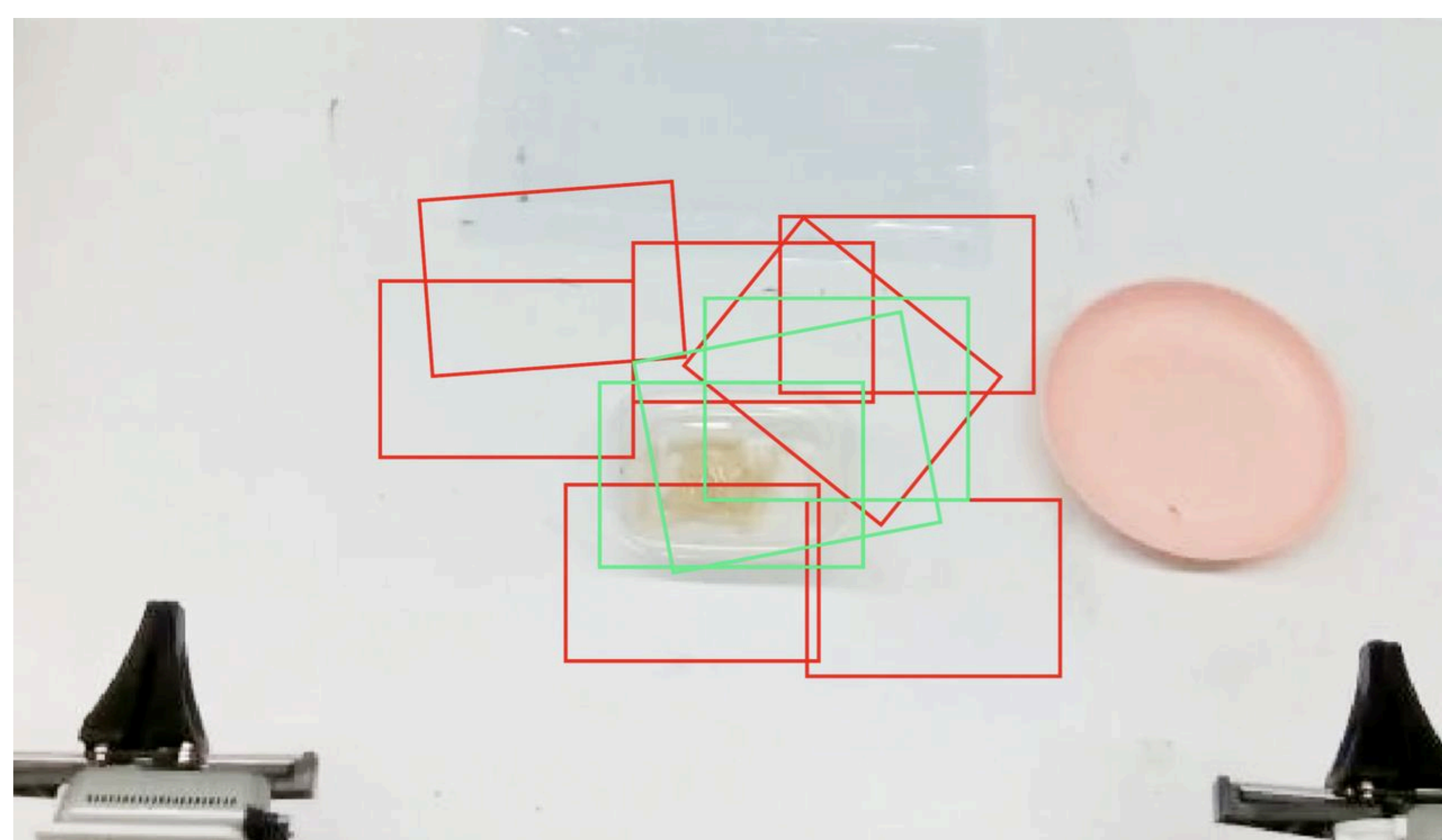
Unlike evaluation modes that require the submission of model weights or Docker images, the RoboChallenge platform adopts a user-side inference architecture. In this mode, the model runs on the local computing resources of the participant and interacts with the evaluation platform's entity through a standardized online API.



Users do not need to deal with complex environment configuration migration issues; they simply receive observation data with precise timestamps via the API, including RGB images, depth maps, and proprioception data, as well as action commands.

■ 5.2 Eliminating Human Bias: Visual Input Matching Method

In real machine evaluations, the largest variable often comes from the operator's arrangement of the test scene. Research has shown that experienced operators may unconsciously place objects in locations where models are more likely to succeed (so-called "sweet spots"), leading to biased and distorted evaluation results. To eliminate this systemic bias, RoboChallenge introduces the Visual Task Reproduction method. During each test reset phase, the system extracts a reference image frame from the training data and overlays it semi-transparently on the operator's real-time preview screen. Test personnel must continuously adjust the object positions until the real-time image visually aligns closely with the reference image. This approach ensures that the initial state distribution of each test remains consistent with the training data, significantly enhancing the stability and fairness of the evaluations.



■ 5.3 Multi-Dimensional Quantitative Scoring Standards

To measure model performance more precisely, RoboChallenge adopts a dual evaluation system of success rate (Success Rate) and progress score (Progress Score).

- **Success Rate:** An indicator measuring whether the overall task is completed.
- **Progress Score:** An indicator measuring the degree of advancement towards the success goal during task execution. RoboChallenge breaks each task down into several execution stages, and points are awarded for completing each stage. Even if the final task fails, the partial progress achieved in the intermediate stages will still be quantified and recorded.
- **Retry Penalty:** To address the common instability in actual operations, the system introduces a retry deduction mechanism. If the entity fails an attempt in a certain stage (e.g., dropping the grasp), and retries, 0.5 points will be deducted for each retry. When the score for a stage drops to negative, or if the number of consecutive failures exceeds 4, that round of testing will automatically terminate.

■ 5.4 Open Source Real Machine Dataset

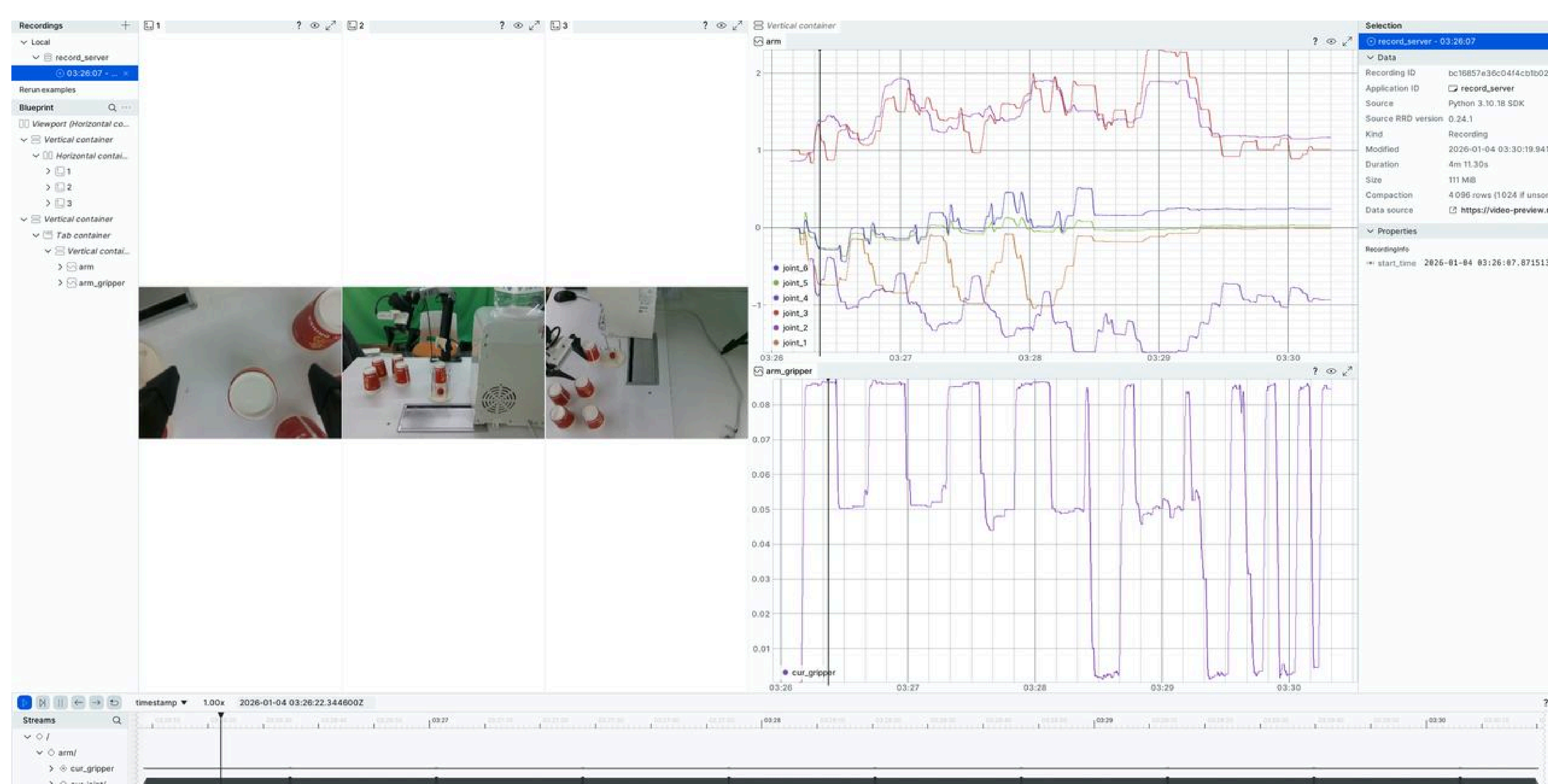
The RoboChallenge platform provides rich real machine demonstration data for each task in the Table30 benchmark, with a single task containing up to 1000 complete trajectories (Episodes).

All datasets are hosted on Hugging Face and are fully open to the community. The data includes not only the original video files and JSON format entity state information but also conversion scripts that support transforming the data into mainstream formats like LeRobot, greatly lowering the barrier for developers.

■ 5.5 Submission Evaluation and Transparency

In terms of submission mechanisms, the platform distinguishes between task-specific and multitask model settings.

Additionally, to promote community interaction, RoboChallenge will publicly share participants' evaluation recordings and machine logs (in RRD format) on its official website, allowing researchers to review, analyze, and learn from failure cases, thereby advancing algorithmic progress.



6. RoboChallenge Partner Ecosystem

■ 6.1 The Evolution of Community Development

After Dexmal and Hugging Face jointly initiated RoboChallenge, it quickly resonated strongly within the industry and attracted widespread participation. Partners including BAAI, Agibot, Qwen, Galaxea, X Square Robot, Tsinghua University, Xi'an Jiaotong University, and GOSIM have joined forces to promote ecosystem development, officially establishing the RoboChallenge Committee on November 20, 2025. This marks a new phase of standardized open collaboration in the evaluation of embodied intelligence, injecting new momentum into the implementation and iteration of embodied intelligence technology through an "open community, co-creation" model. The establishment of the RoboChallenge Committee aims to upgrade RoboChallenge to an industry-level public infrastructure, promoting the transition of embodied intelligence evaluation from "dispersed experiments" to "consensus building" through standardized and regular operational mechanisms. In the future, the committee will work with industry, academia, and the open-source community to build a transparent, efficient, and trustworthy evaluation ecosystem for embodied intelligence, facilitating the accelerated formation and implementation of industry evaluation standards.

■ 6.2 The Launch of the Committee and Organizational Structure

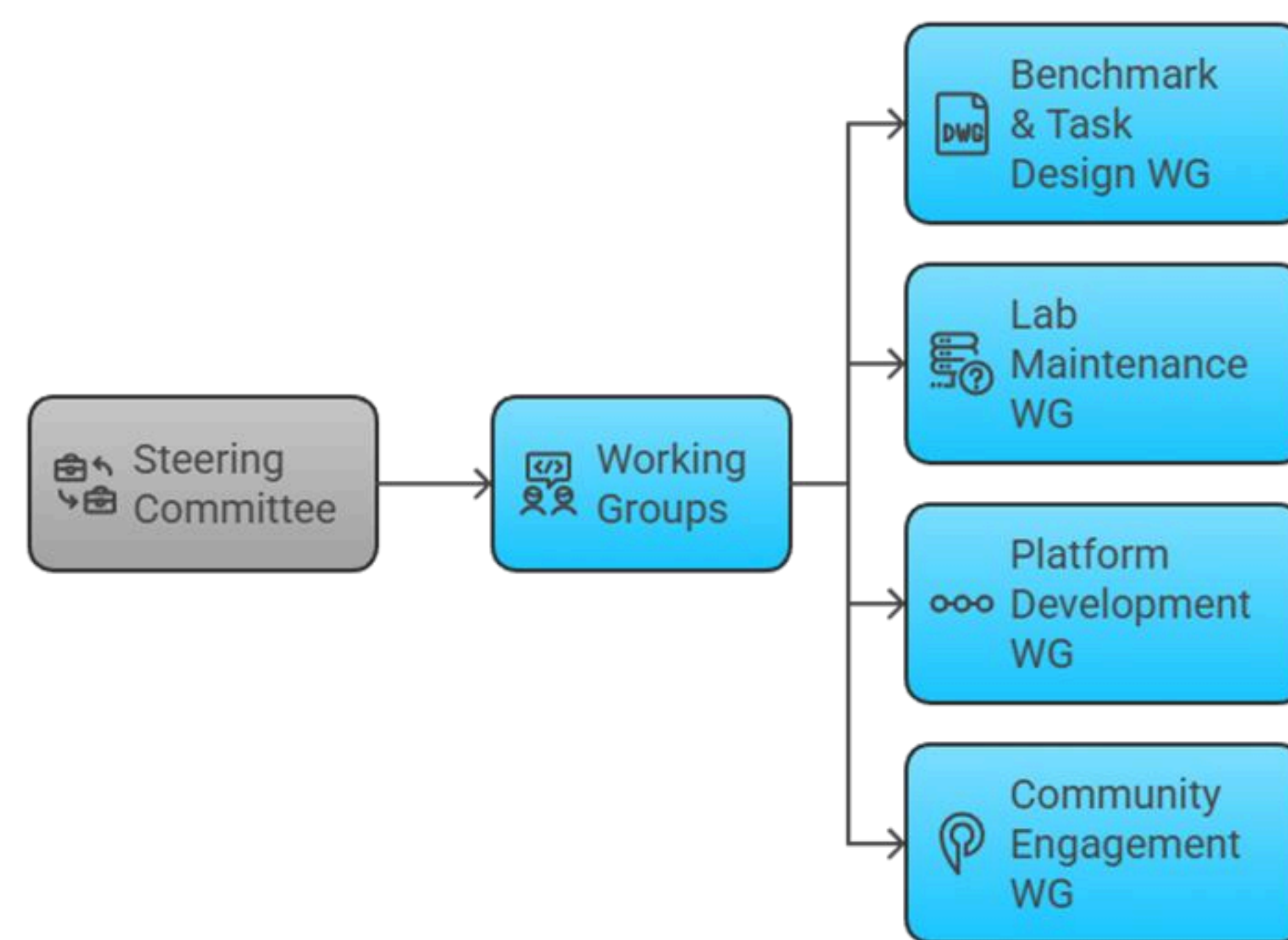
The established RoboChallenge Committee is adopting a dual-layer structure of a Steering Committee and Working Groups to ensure efficient execution and transparent governance. The committee serves as the decision-making hub, adhering to the principles of "openness, collaboration and consensus" while overseeing the overall strategy. It will have four core working groups that act as the "execution engines" for specific implementations.

- **Benchmark & Task Design Working Group:** Focuses on the "core brain" of the evaluation system. This group is responsible for designing and optimizing benchmarks for embodied intelligence, typical tasks, and supporting datasets.
- **Lab Maintenance Working Group:** Acts as the "hardware steward". This group is responsible for receiving donated testing robots from partners, maintaining the laboratory environment, and conducting standardized tests. By standardizing hardware conditions and testing processes, it ensures that evaluation results from different robots are comparable and reproducible.
- **Platform Development Working Group:** Builds the "digital foundation". This group is responsible for developing and maintaining the robochallenge.ai website, developing API services and data distribution tools, allowing global developers to easily upload test data and obtain evaluation results, forming a closed-loop ecosystem of "testing-

feedback-optimization".

- Community Engagement Working Group: Takes on the role of an "amplifier". By organizing hackathons, academic seminars, and online and offline activities, it attracts more developers, enterprises, and universities to join, promoting technical exchanges and demand matching.

RoboChallenge Organizational Structure



The willingness to join the RoboChallenge working group community continues to rise. As of January 23, 2026, we have officially received membership applications from several organizations, including: Sichuan Embodied Intelligence Robot Training Base, China Mobile (Hangzhou) Information Technology, Zeroth, INFIFORCE, AI² Robotics, Lightwheel, Stella-Robot, x-humanoid, Spirit AI, Robbyant, and GigaAI.

In January 2026, Galaxea donated an R1 Lite to the RoboChallenge committee's real machine laboratory, further enriching the hardware configuration library of RoboChallenge and providing strong hardware support for future real machine evaluations across different entities and scenarios.

As the community continues to grow, the RoboChallenge working group community is gradually forming a collaborative network characterized by high-frequency interactions and substantial co-construction. Community members are continuously engaging in in-depth discussions by online meetings and contributions around core topics such as evaluation task design, indicator systems, and real machine reproduction, collectively advancing the improvement of the embodied intelligence evaluation system and the overall progress of the industry.

■ 6.3 Governance Principles: Driving a Sustainable Future through Consensus

In addition to the specific execution structure, the operational principles of the RoboChallenge Organizing Committee are also noteworthy; its core commitments include three points: testing is open and reproducible, the community is inclusive and non-competitive, and contributions are open-source and traceable. This means that the technical details, test data, and improvement plans of any participant must be made public within the framework to avoid "reinventing the wheel"; the evaluation of embodied intelligence is not a knockout competition but a collaborative space. Through open

governance, the RoboChallenge Organizing Committee hopes to attract more research institutions, enterprises, and startups to join—these entities are often closer to real needs and can inject fresh scenarios into the evaluation system. For example, companies producing home service robots may focus on the evaluation dimension of "multi-language command response", which will be incorporated by the working groups to ultimately benefit the entire industry.

■ 6.4 Community Feedback and Suggestions for the Future Development of RoboChallenge

In the past three months, RoboChallenge has continuously interacted with researchers, developers, and enterprise partners, accumulating a wealth of constructive suggestions for future development. Based on feedback from the RoboChallenge Organizing Committee, the four working groups, and the broader community partners, we have summarized the following key directions and specific suggestions:

1) Enhance the coverage and systematization of Zero-Shot capability evaluation.

The RoboChallenge community has repeatedly mentioned that current real robot evaluations often focus on pre-trained or specifically fine-tuned tasks, but the evaluation system for "unseen tasks / Zero-Shot execution capability" remains inadequate.

- Establish Zero-Shot capability standards and task hierarchy indicators;
- Develop a dedicated Zero-Shot benchmark for instruction understanding + environmental perception + action generation;
- Compare the performance differences of different models under Zero-Shot conditions and publicly release reproducible evaluation scripts and analysis reports.

2) Build a truly universal multitask evaluation system.

The community generally hopes that RoboChallenge can support systematic evaluation of multi-task general models, not just single-task rankings, but overall multi-task coordination capabilities.

- Design cross-task combination schemes, such as labeling by task type (picking/assembly/inference/complex sequential actions, etc.); combine multiple sub-tasks into a complex composite task.
- Launch a "Universal Strategy Challenge" to encourage submissions of models that maintain performance across multiple tasks;
- Establish a unified scoring mechanism that includes accuracy, continuity, execution efficiency, and other multi-dimensional evaluation indicators.

3) Visualize the comparison of real environment vs. simulation evaluation effects.

An important demand from partners is to understand the performance gap of the same model in simulators versus real environments, which is crucial for understanding model generalization capabilities.

- Increase the testing task sets for both simulation and real machines in a 1:1 correspondence. For example, if Table30 already exists, consider generating a simulation environment for Table30.
- If there is a set of simulation tasks, it should correspondingly support a task set for real machines.
- Regularly publish comparison curve graphs, such as "simulation scores vs. RoboChallenge real environment scores" visualizing the differences between the two using a unified task set and scoring standards;
- Provide basic visualization templates and customized reporting functions to facilitate the community in generating comparison curves;
- Continuously accumulate such horizontal comparison data to provide a quantifiable basis for real transfer difficulty in future research.

4) Increase support for co-hosting competitions and industry collaborations.

Community and institutional feedback indicates a desire for RoboChallenge to become a more influential competition platform in the field of embodied AI, such as collaborating with top conferences (ICRA / RSS / NeurIPS Robotics Track) or iconic events.

- Proactively invite universities, research institutes, and industry alliances to become co-hosts of competitions;
- Collaborate with simulation platforms and robot suppliers to provide official benchmark awards and real machine testing support;
- Explore incorporating RoboChallenge evaluations into the official scoring sub-tracks of some mainstream international robotics competitions.

5) Continuously expand the scale and diversity of benchmarks.

The existing Table30 of RoboChallenge is a good start, but the community calls for more dimensions, more complex scenarios, and more challenging task sets.

- Introduce higher-level task benchmarks, such as complex assembly, dynamic scene actions, and multi-person collaboration tasks.
- Collaborate with the community to build "Bench30+ / Bench100 series task sets", including tasks submitted and reviewed by the community;
- Regularly publish benchmark library updates and leaderboards to enhance the activity of the dataset ecosystem.

6) Increase the variety of robot types and compatible interfaces.

Partner feedback indicates that there are many types of real robots, and adapting to different models often requires repetitive work.

- Support more bodies, dexterous hands, and other robotic devices to connect to the RoboChallenge evaluation framework;
- Prioritize models that are in high demand from the community and have broad industrial applications.
- Launch a "Model Adaptation Co-construction Plan" to encourage hardware manufacturers to submit drivers and adaptation codes for official support.

7) Increase the dimensionality of evaluation scenarios and real-world complexity.

Real-world embodied tasks often come with challenges such as uncertainty, dynamic changes, and confined spaces. The community suggests that benchmarks should not be limited to static desktop sets.

- Release multi-scenario benchmarks, such as kitchens, homes, warehouses, and component assembly lines;
- Introduce some dynamic elements (movable obstacles, changing target positions, etc.).

8) Increase the transparency of the queuing mechanism.

From the user's perspective, users do not know how long they need to wait for evaluation, and many people cannot guarantee that every user can be tested when submitting tests. Suggestions include:

- Transparency of the queue system: set up a real-time, transparent queue list (or waitlist), or display the current number of pending tasks in the background. Automatically trigger tests according to submission order to reduce communication costs and make the evaluation process more in line with the collaborative habits of the international open-source community.
- Submission quota management: to balance server load and resource fairness, it is recommended to limit the number of submissions per account per day. This can effectively control the total task volume and guide participating teams to fully validate locally before submitting, improving overall evaluation efficiency.

9) Build a rigorous generalization and "blind testing" system.

From the perspective of scientific rigor in evaluation, to prevent models from "overfitting" to specific environments or obtaining high scores through "cheating", a multi-dimensional dynamic assessment mechanism can be established. Suggestions include:

- Comprehensive generalization assessment: establish a three-level generalization testing standard from objects, positions to scenes. By randomly changing the placement of target objects, introducing unseen similar objects, and altering background lighting or textures, comprehensively test whether the model truly understands "semantics" or merely remembers "pixels".
- Strict confidentiality of hardware configurations: implement a "blind testing" mechanism for evaluation environment parameters, keeping key data such as camera pitch angles and installation heights confidential. This forces models to rely on visual perception and adaptability to complete tasks rather than using known parameter priors for hard-coded cheating, ensuring the authenticity and fairness of evaluation results.

7. Platform Operation Review

RoboChallenge, as a pioneering real-machine evaluation platform, has undergone a leap from "prototype validation" to "scaled services" since its launch. Facing the complexities of the real physical world and the surging industry demands, we confront challenges head-on, continuously refining and iterating to strengthen the scientific rigor and authority of our evaluation system.

■ 7.1 Capacity Bottleneck Issues

At the initial stage of the platform launch, the system was designed with a daily capacity threshold of 800 real device tests (Rollouts). However, the actual demand for evaluations grew at a pace far exceeding expectations, especially in December, where the daily demand for real device tests (Rollouts) stabilized between 1000 and 1500, showing explosive growth. In the face of a severe supply-demand gap, the operations team successfully increased the system's measured capacity by 50% through various strategies, including optimizing scheduling algorithms and extending service hours.

■ 7.2 Testing Consistency Challenges

Although RoboChallenge established a standardized system at the outset, including task design, data standards, and scoring rules, the "uncontrolled nature" of the real physical environment remains the biggest variable. We must acknowledge that absolute consistency is an ideal state that is difficult to achieve in an uncontrolled physical world.

Actual measurement data shows that even in strictly controlled laboratory environments, slight fluctuations in lighting or differences in the operator's clothing color can lead to score variations for the same model across two submissions. Currently, we have not been able to completely eliminate these physical "noise" factors, but these real-world problem cases have not only helped us better understand the model's sensitivity to various interference factors but also accumulated valuable experience for the scientific design of future evaluation sets and the rigor of evaluations.

■ 7.3 Secondary Review of Scoring

In response to potential subjective biases in manual evaluations, we realized that a single scoring sheet could no longer meet the requirements for rigor. Therefore, the platform has built an evaluation knowledge base containing numerous standard cases and introduced a secondary review mechanism and appeal arbitration channels. Although we cannot guarantee a 100% initial judgment accuracy rate, this error correction process has significantly reduced statistical bias in results, establishing a scoring principle of correcting errors and ensuring traceability, helping RoboChallenge's evaluation results to be more rigorous.

■ 7.4 Trust Mechanism and Open Source Appeal

Given the user-side inference architecture adopted by the RoboChallenge platform, as a third party, the platform objectively cannot verify the authenticity of users' proprietary model architectures from a technical standpoint, nor can it completely avoid targeted tuning issues. This is a limitation faced by remote real-machine evaluations. Therefore, RoboChallenge sincerely calls on the community to embrace open source and transparency. We always adhere to a neutral service principle, but only through the disclosure of code and weights, combined with empirical real-machine evaluations, can we completely break the suspicion.

■ 7.5 Remote Interaction Directly Affects Capacity Limits

RoboChallenge adopts a user-side inference architecture, which greatly reduces the user's environmental migration costs and ensures the safety of model assets, but also brings about significant efficiency losses. In actual operation, the overall efficiency of real-machine evaluations no longer solely depends on the robot's execution speed but is deeply coupled with network transmission delays and user-side model inference time. Even millisecond-level network fluctuations or insufficient computing power on the user side can directly translate into the robot's "idle waiting", thereby extending the time occupied by a single task and restricting the platform's overall daily throughput. How to further compress communication delays and improve the collaborative efficiency of remote inference will be the core issue we tackle in the next phase through technical breakthroughs.

8. Participants' Experience and Insights

■ 8.1 Participant Practical Review - Insights from Community Volunteers' Real Testing Experience

Community volunteer from TAO-DualArmVLA model team shared his practical experience in RoboChallenge evaluations.

1. Data Engineering: High-quality "fuel" is half the success. Before model training, I invested a lot of effort in raw data governance, and practical tests proved that this investment effectively improved the scores of real-machine evaluations.

- **Deep observation of raw data:** Through visualization methods, I conducted intuitive analyses of video and sensor data, accurately grasping the distribution patterns, scale ranges, and boundaries of outliers, providing a basis for subsequent processing.
- **Signal cleaning and discretization:** For the continuous waveform signals collected by the gripper sensor in the raw data (which usually contain unnecessary noise and interference), to enhance training effectiveness, I converted them into square wave signals representing the "open/close" states. This processing significantly improved the model's stability in judging the state of the end effector.

2. Strategy Selection: From Points to Planes, Finding Breakthroughs.

In the face of a multi-task concurrent scoring system, blindly pursuing "full bloom" is often inefficient. I adopted a point-to-plane breakthrough strategy:

- **Focusing on high-scoring tasks (Critical Path Method):** Prioritize establishing baselines for tasks with high scoring weights that are easy to break through, concentrating resources to clear critical paths.
- **Cross-comparison and strategy evaluation:** Based on single-point breakthroughs, conduct cross-comparison and evaluation of strategies to quickly filter out general modules with transfer value.

3. Model Training: Scientific Experiments, Iterative Optimization.

We view model training as a rigorous scientific experiment process, abandoning the random trial-and-error model:

- **Systematic experimental documentation:** Use tools to detail every experiment's hyperparameters, code versions, dataset versions, and corresponding performance metrics, achieving full-link traceability.
- **Strict variable control:** Adhere to the control variable method for parameter tuning, changing only 1-2 key parameters at a time. This approach allows for clear attribution of performance changes and ultimately forms a clear experimental report, ensuring the correctness of the direction for continuous optimization.

■ 8.2 Completed Participants - Insights from the Spirit AI Team's Evaluation Experience

During the development and evaluation process of Spirit-v1.5, we conducted a systematic analysis of the model's generalization bottlenecks in complex physical environments and achieved the following key evolutions in algorithm architecture and control strategies:

- **Decoupling Strategy for Fine Operations:** In the evaluations, we observed that the model tends to overfit when executing high-precision tasks, excessively relying on the feedback state of the robotic arm while neglecting real-time visual guidance, leading to deviations in the end effector's pose at critical moments. To address this, we innovatively proposed the Mask Z strategy: masking the height (Z-axis) information in the input state, forcing the model to infer object height through vision in real-time. This change significantly enhanced the model's perceptual generalization across different desktop environments, effectively reducing localization errors.
- **Dynamic Stability Optimization at Extreme Positions:** For hardware platforms like Aloha that tend to exhibit end effector jitter at the limits of their workspace, we introduced an action smoothing mechanism on the inference side. By performing temporal smoothing on the output Action Chunk, we ensured the stability of the robotic arm's operation at extreme positions, greatly improving the coherence of task completion.
- **The Impact of Visual Resolution on Spatial Awareness:** The leaderboard results revealed that the model still experiences empty grasping phenomena due to insufficient spatial awareness when facing small object grabs, primarily limited by the input resolution of the visual encoder during the pre-training phase. We plan to introduce a higher resolution visual encoder to address this issue in the future.

We thank the organizers for building a high-level embodied intelligence evaluation platform that allows us to validate the generality of Spirit-v1.5 in a fair and open environment. Special thanks to the official team for providing technical support during model deployment and real-machine debugging; your feedback is a significant driving force for our model iterations. Looking ahead, we look forward to working with RoboChallenge to promote and witness the GPT-3 moment of embodied intelligence.

■ 8.3 Completed Participants - Insights from the X Square Robot Evaluation Experience

During the completion of the full process evaluation of RoboChallenge Table30, the Wall-oss team from X Square not only validated the model's boundary capabilities in multi-dimensional tasks but also accumulated valuable engineering experience through interactions with the real physical world. Here are our feedback from the practical process:

- Establishing a Mapping Relationship Between "Open-Loop Indicators and Closed-Loop Measurements". Given the scarcity of real-machine evaluation resources and time costs, directly relying on full real-machine testing is inefficient. Our team recommends prioritizing sufficient open-loop evaluations, analyzing the cross-correlations between Loss curves, open-loop action trajectories, and final real-machine success rates to summarize a reliable predictive pattern. Utilizing this pattern to guide model selection (Checkpoint) can significantly reduce ineffective real-machine submissions and achieve a multiplication of evaluation efficiency.
- Inference Parameters Need to Be Fine-Tuned for "Remote Real Machines". There is no universal solution that fits all parameters. For RoboChallenge's cross-entity (different models) and cross-task characteristics, action execution length, control frequency, and end strategy inference parameters must be specially tuned. Additionally, since RoboChallenge adopts a remote evaluation mode, network delays and fluctuations are significant interference factors.
- In multi-step long-range tasks, a phenomenon of "state confusion" often occurs, where different stages have visually similar inputs (such as the stationary state before and after an object is picked up). Relying solely on the current frame often leads to decision loops. Practical experience has shown that introducing a strong temporal state transition module is crucial; it helps the model utilize historical context information to discern the current state, significantly enhancing task execution coherence and fluency.
- It is worth noting for subsequent participating teams that RoboChallenge has now implemented stricter submission standards. The total score calculation for Table30 is strictly based on the complete submission records of a single account. We recommend that teams manage accounts uniformly and concentrate resources to complete systematic testing of all tasks to ensure that final scores can be accurately recorded on the Leaderboard.

The RoboChallenge Table30 task set covers core capabilities of embodied intelligence such as fine visual perception, temporal logic understanding, and multi-entity adaptation. We thank the organizing team for their efforts in building a standardized evaluation system. During the process, we also worked with the organizers to solve the end control issues in the cross-entity generalization of robots. In the future, we hope to continue collaborating with RoboChallenge to promote the evolution of evaluation standards and look forward to witnessing embodied intelligence move towards generalization.

■ 8.4 Ongoing Participants - Insights from the GigaBrain Team

During the first-round exploratory evaluation of GigaBrain, a fully self-developed and end-to-end open-source embodied foundation model, on the RoboChallenge Table30 tasks, GigaBrain team summarized several key lessons from a hands-on, real-world perspective. These insights are intended to serve as practical references for the community in embodied model training and inference optimization:

- Pretraining coverage of the base model defines the upper bound of capability. In the early stages of evaluation, we observed that task difficulty and model performance are not always positively correlated. Some intuitively “simple” tasks achieved success rates of only around 80% due to mismatches in training data distribution, while certain “complex” tasks exhibited strong robustness when they aligned well with the base model’s pretraining domains (e.g., spatial representations or specific manipulation semantics). This indicates that the breadth of pretraining data fundamentally determines the generalization threshold of downstream tasks, and post-training alone cannot fully compensate for missing cognition in long-tail distributions. The world-model-centric technical path pioneered and practiced by GigaAI addresses this generalization challenge at its root. Its effectiveness is not only demonstrated on RoboChallenge tasks, but also points toward the genuine realization of Physical AGI.
- Establish a mapping between “open-loop metrics” and “closed-loop real-world performance.” Given the scarcity and high time cost of real-robot evaluation resources, relying solely on large-scale physical testing is inefficient. We recommend conducting sufficiently rich open-loop evaluations first. By analyzing the cross-correlations among loss curves, open-loop action trajectories, and final real-robot success rates, we derived a set of reliable predictive patterns. These patterns can guide checkpoint selection and significantly reduce ineffective real-robot submissions, effectively doubling evaluation efficiency.
- Inference parameters must be finely adapted for “remote real-robot” evaluation. There is no one-size-fits-all parameter configuration. To accommodate RoboChallenge’s cross-embodiment (different robot platforms) and cross-task characteristics, inference parameters—such as action horizon, control frequency, and end-effector strategies—must be specifically tuned. Moreover, because RoboChallenge adopts a remote evaluation paradigm, network latency and jitter are non-negligible factors that must be explicitly accounted for.
- Effective use of temporal information yields disproportionate gains. In multi-step, long-horizon tasks, “state ambiguity” frequently arises when visual inputs across different phases appear highly similar (e.g., before and after an object is picked up while remaining visually static). Relying solely on the current frame often leads to decision dead loops. To address this, we introduced a strong temporal state transition module, enabling the model to leverage historical context to disambiguate the current state. This significantly improves execution continuity and overall task smoothness.

RoboChallenge is the world's first large-scale real-machine evaluation platform for embodied intelligence, with significant practical guidance implications. We look forward to collaborating closely with RoboChallenge to co-create and elevate benchmarks for physical intelligence—and to jointly witness the arrival of the Physical AGI moment.

Acknowledgments:

The Dexamal team provided real-machine cluster resources and long-term operational manpower support for the RoboChallenge platform, making large-scale, standardized real-machine evaluations possible.

The RoboChallenge committee has gathered core forces from industry, academia, and research, actively participating in discussions and continuously contributing insights on key topics such as evaluation benchmark system design, evaluation data analysis and review, future platform evolution paths, and industry development trends, jointly promoting the maturity and evolution of the embodied intelligence evaluation ecosystem.

Report Editorial Committee

Yang Chen, Xue Sun, Haoqiang Fan, Teng Ma, Farman Huang, Enwen Zhang, Kai Liu, Junliang Guo, Pushi Zhang, Pengwei Wang, Tiezhen Wang, Nadia Jiang, Xiaomeng He, Yikuan Chen, Yang Wang, Zheng Zhu, Congcong Yu.