RoboChallenge

年度报告

2025 Q4 ~ 2026 Q1

2025 Annual Report

Published by the RoboChallenge Committee

RoboChallenge 年度报告

2025 Q4 – 2026 Q1

随着大语言模型（LLM）与视觉语言模型（VLM）的爆发式突破，人工智能已在数字空间展现出惊人的认知与推理能力。 如今，这股技术浪潮正加速向物理世界外溢，推动机器人从单一功能的自动化设备，向具备通用理解与决策能力的具身智能（Embodied AI）进化。然而，要实现从「实验室智能」向「现实世界智能」的跨越，行业仍需克服真实环境下的验证难、测试条件非标、公开平台缺失等挑战。构建一个开放、公正、可复现的「真实考场」，已成为行业迈向通用化的必经之路。

RoboChallenge 正是为解决这一核心命题而生。作为由原力灵机 Dexmal 与 Hugging Face 联合推出的全球首个具身智能大规模真机评测平台，RoboChallenge 致力于通过科学的评测体系，为视觉-语言-动作模型（VLAs）在机器人的实际应用提供更加可靠和可比较的评测标准。自 2025 年 10 月 15 日上线以来，平台已成功部署了包含 UR5、Franka、ARX5、ALOHA 等主流机型在内的 20 台真机测试集群，并开源了涵盖 9 大类、30 个标准化桌面任务的 Table30 数据集。

为进一步凝聚行业力量，加速具身智能真机评测标准的规范化进程，原力灵机 Dexmal 与 Hugging Face 联合智源研究院、智元机器人、Qwen、星海图、自变量、清华大学、西安交通大学及 GOSIM 等单位，于 2025 年 11 月 20 日正式成立 RoboChallenge 组委会，旨在通过产学研深度协同以广泛的行业共识定义具身智能真机评测标准。

RoboChallenge 一经上线便迅速引起了全球具身智能社区的积极响应。由社区和个人提测的 π0 与 π0.5、RDT-1B、CogACT 及 OpenVLA-OFT 等开源模型已成功上榜。目前，千寻智能与自变量团队已完成完整的 Table30 评测并上榜。与此同时，更多行业机构正陆续加入真机实测：极佳视界、智源研究院、中移杭研、星海图、地平线等单位的模型也均在测试中。这种跨越国界、汇聚顶尖研究院所、科技巨头、独角兽及开源社区的参与趋势，代表着拥抱真机实测、标准化评测已成为具身智能领域的行业共识。

基于此，本报告深度分析了 RoboChallenge 过往数万次真机实测数据，为您呈现现阶段 VLA 模型在物理环境下的真实表现。然而，这仅仅是 RoboChallenge 推动具身智能真机评测标准化的序章。未来，我们将持续引入更多机型的本体，拓展多元化与真实工业场景评测集，推出更具挑战性的真机评测任务，并探索分布式真机评测机制。我们的愿景是与社区一道，通过最真实的具身智能真机评测平台，助力具身智能在真实物理环境中创造价值。

一、 核心发现与亮点观察

1. **评测热度呈指数级增长，真机验证成为行业刚需。**在过去的三个月中，RoboChallenge 平台注册用户数与评测提交量均呈现出显著的指数级增长趋势，评测热度的持续走高也标志着 RoboChallenge 已成为检验具身智能模型能力的重要平台。

2. **叠碗和物体移入盒子已成为“Hello World”级任务。**数据显示，大多数的提交用户都将这两项任务作为 VLA 模型评测的首选测试任务，且成功率较高。

3. **整理纸杯、制作三明治等复杂任务依然是待攻破的难题。**涉及多步骤操作和精细操控的任务，对于当前参测模型而言依然难度偏高，成功率长期处于低位，部分任务甚至接近 0。

4. **榜首模型成功率约 50%，依然有进步空间。**当前参测模型在 Table30 评测集上的最高成功率仅约 50%，仍有提升空间，同时也体现了 Table30 任务集的挑战性。

5. **VLA 模型仍在攻克人类的本能级操作。**实测数据显示，参测模型虽具备较强的指令语义理解能力（呈现移动趋势），但在精细操作任务中成功率不足 15%。这种现象在 RoboChallenge 平台上沉淀了大量真机失败数据，这份公开的“错题集”可作为模型迭代优化的关键参考。

二、 数据洞察

2.1 平台用户数据

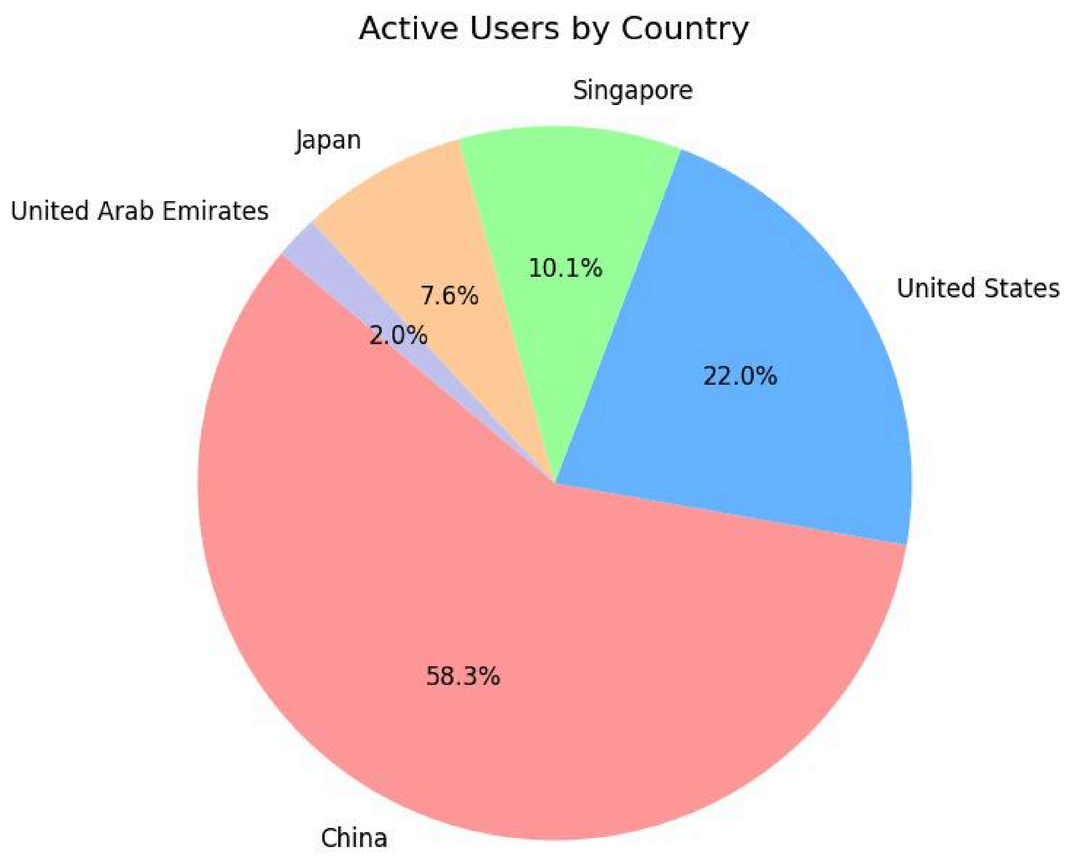
自 RoboChallenge 平台上线以来，社区参与热情持续高涨，平台在这段时间承载了大规模的评测需求。

- 高提测转化率：平台累计核发提测资格 209 个。其中，已有 82 位开发者成功跑通了本地推理环境并提交了评测请求。从申请到实际提测的转化率达 39.2%，反映了具身智能领域对标准化真机评测的刚需。

- 大量评测数据积累：平台累计执行的真机测试（Rollouts）总数已突破 4 万次（41969 次）。平台在任务提测、真机执行、日志记录、结果发布等环节已形成评测闭环，也为后续深入分析提供了数据基础。

- 高并发吞吐能力：平台单日模型提交评测达到 181 次（Runs），单日真机测试峰值为 834 次（Rollouts），此数据验证了 RoboChallenge 平台架构在高负载环境下的稳定性与可靠性。

2.2 网站访问数据分析



活跃用户地域分布显示出 RoboChallenge 正在形成的国际化生态。尽管中国开发者目前作为核心力量占 58.3%，但来自美国（22.0%）、新加坡（10.1%）、日本（7.6%）和阿联酋（2.0%）的开发者也在持续涌入，来自不同国家的开发者正在同一套标准化真机评测平台上同场竞技。

2.3 开源平台数据汇总和分析

自 RoboChallenge 发布之初，平台即坚持“评测与数据同步开源”的原则，在 GitHub 与 Hugging Face 两大主流开源平台上同步开放了与 RoboChallenge 相关的数据集、任务定义与配套资源，降低研究者与开发者的使用门槛，推动真实机器人评测的可复现与可对比。

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

在 Hugging Face 平台（https://huggingface.co/RoboChallenge）上，RoboChallenge 正式发布了 Table30 全量数据集，涵盖 30 个标准化桌面操作任务，支持模型训练、离线评估与结果复现。截至目前，Table30 数据集累计下载量已达 17K 次（近一个月下载 7k 次），显示出社区对统一真实机器人任务基准的持续关注与实际使用需求。该数据集已被广泛用于模型能力对比、方法消融实验以及“从仿真到真机”的迁移研究中，逐步成为具身智能领域的基础参考资源之一。

与此同时，在 GitHub（https://github.com/RoboChallenge）上，RoboChallenge 开源仓库集中发布了平台相关的评测代码、任务配置示例与使用说明，方便研究者快速接入真实机器人评测流程。并持续收到来自社区的 issue 与改进建议，反映出 RoboChallenge 正在形成以“真实评测 + 开源协作”为核心的早期开发者与研究者社区。

通过在 GitHub 与 Hugging Face 上的双平台协同开源，RoboChallenge 不仅实现了数据、代码与评测流程的透明化，也为后续更多模型、任务与硬件平台的接入打下了社区基础。这一策略使 RoboChallenge 能够在平台规模尚处于早期阶段时，就提前进入研究者与开发者的工作流之中，而非停留在展示层面。

三、 榜单深度解读

RoboChallenge 作为全球首个具身智能大规模真机评测平台，其榜单已成为衡量模型真机实测表现的重要参照。自发布以来，RoboChallenge 榜单凭借统一的测试标准和公开透明的评测机制，迅速在开源社区与产研界引起广泛影响。榜单结果不仅被众多开发者转发，也获得多家头部模型团队的认可，逐渐发展成为验证算法迭代效果与模型真实性能的重要评测平台。

在深入分析 RoboChallenge 榜单之前，首先需要了解其入围门槛和公开透明机制。
RoboChallenge 榜单并非对分数的简单罗列，而是建立在以下三个关键核心规则之上。

1. 总榜准入门槛：不同于仅针对单一技能（如抓取）进行测试的榜单，RoboChallenge 采用更加全面的评测标准。仅有完整完成 Table30 评测集中 30 个任务的模型，才有资格进入总榜。这意味着，每一个上榜模型都经历了从基础刚体抓取到复杂长程、柔性物体操作的系统性考验。

2. 公开透明的评测机制：RoboChallenge 致力于构建公开透明的评测生态。参测者的评测录像、机器日志均在平台上公开发布。无论是成功率高达 100% 的完美操作，还是成功率为 0% 的失败尝试，所有数据均向全球开发者和用户开放查阅。

3. 系统性消除人为偏差：为降低操作员主观因素对评测结果的影响，RoboChallenge 引入了视觉输入匹配方法。通过在操作端叠加“半透明参考图像”，要求每次测试前将物体摆放位置调整至与训练数据中的参考场景高度重合。该机制从源头上消除了因操作员主观摆放（如无意识地将物体放在更易抓取的“甜点区域”）所带来的系统性偏差，确保不同模型在尽可能一致的初始条件下开展公平评测。

3.1 总榜

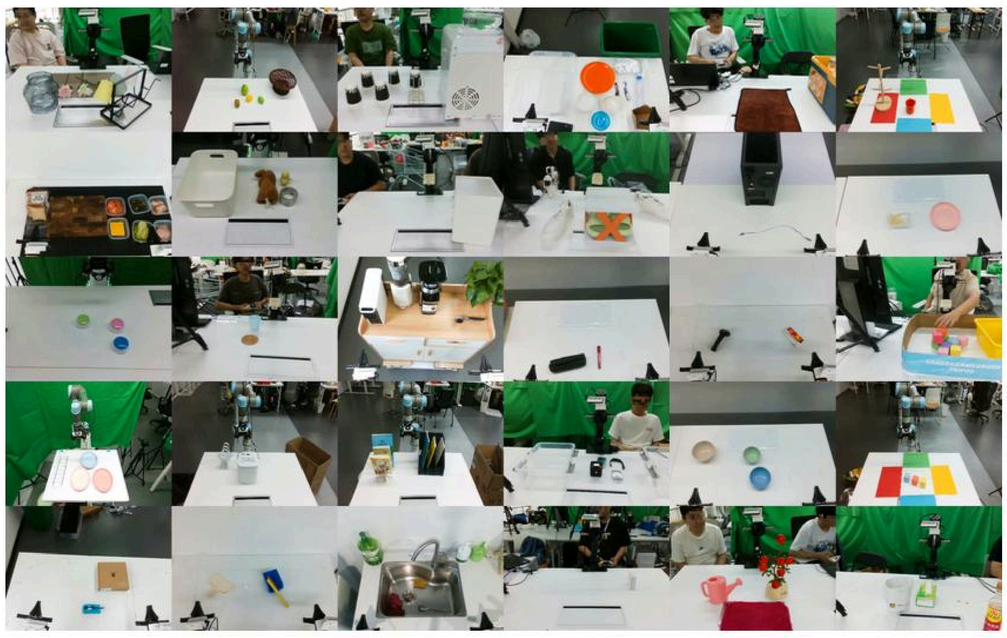
RoboChallenge 首页总榜是按照成功率 (Success Rate) 从高到低排序，我们认为这一指标最能直接反映模型解决实际问题的核心能力。

为了更精细地呈现模型表现，RoboChallenge 采用成功率与过程分的双重评价体系，并制定了严格的定量计算标准。为减小单次运行的偶然性误差，每个模型需针对 Table30 中的每一个任务执行 10 次真机测试（Rollouts）。

- 成功率 (Success Rate)：
  - 单任务计算：在 10 次执行中，成功次数所占比例。例如成功 5 次，则该任务成功率为 50%。
  - 总榜计算：对全部 30 个任务的成功率取平均值。
- 过程分 (Progress Score)：
  - 单任务计算：每次执行满分为 10 分（每次重试扣 0.5 分），10 次执行的得分累加，即为该任务最终得分（满分 100 分）。
  - 总榜计算：对全部 30 个任务的任务得分（各 0–100 分）取平均值。

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-1B/zsz	28.84	**15.00%**
7	cogact/hsk	21.83	**11.67%**
8	pi0_generalist/wyf	20.22	**9.00%**

首页总榜（仅显示 Top8）截图日期：2026.1.23



任务缩略图

数据观察

- Table30 测试的挑战性：榜首模型的平均成功率为 51%。这一数据客观揭示了真机评测的严苛性，即便是本次评测中表现最好的模型，在面对 Table30 所涵盖的刚体、软体及长程等综合任务时，其端到端执行成功率仍仅维持在约 50% 的水平。

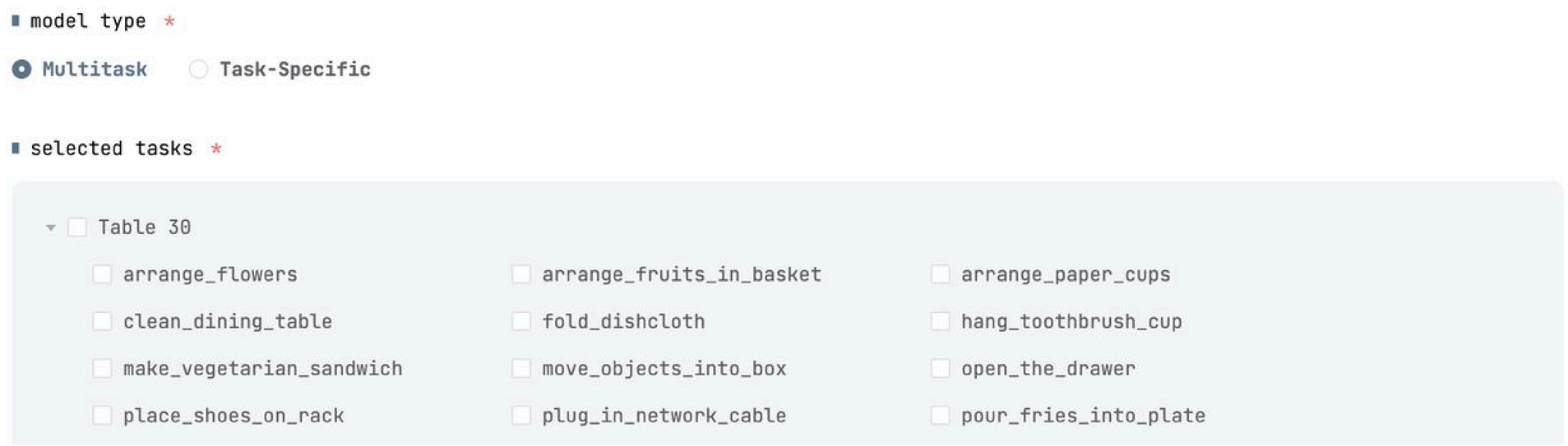
- 过程分揭示潜在能力：数据分析显示，各模型的得分（Score）普遍高于其成功率（例如 Rank 2 模型：Score 61.84 > SR 42.67%）。参测模型在失败任务中并非完全“无所作为”，而是在相当多的情形下已完成超半数的关键步骤。

- 模型能力还有提升空间：当前 51% 的最高成功率意味着具身智能模型在真机评测任务上的整体能力仍有巨大的提升空间。我们相信，行业内仍潜藏着大量尚未参与评测的模型。RoboChallenge 诚挚邀请全球开发者参与评测，在这一公平、标准化的平台上展示更强的模型能力，共同不断刷新最高成功率。

3.2 单/多任务模型榜单

RoboChallenge 平台在评测提交流程中区分了单任务模型与多任务模型，解释如下：

- 单任务模型（Task-Specific）：仅针对单一特定任务进行微调或训练的模型。
- 多任务模型（Multitask）：指单一模型能处理 Table30 中的 2-30 个不同任务的模型。



为便于用户和研究者开展针对性分析，RoboChallenge 官网的 Leaderboard 页面提供了筛选查看的功能，支持用户分别查看单任务模型与多任务模型的排名结果。

Rank	Model/User	Is multitask	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 Zhang	✗	55.30	35.33%
4	pi0/rc_baseline	✗	46.41	28.33%
5	RDT-1B/zsz	✗	28.84	15.00%
6	cogact/hsk	✗	21.83	11.67%
7	openvla-oft/gkf	✗	8.66	5.00%

单任务榜单

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

多任务榜单

对比同一基座模型在单任务与多任务设定下的表现：

- **pi0.5：** 从单任务模型的 42.67% 下滑至多任务模型（pi05_generalist）的 17.67%，成功率下降约 25 个百分点。
- **pi0：** 从单任务模型的 28.33% 下滑至多任务模型（pi0_generalist）的 9.00%，成功率下降近 20 个百分点。

开发者邀请：挑战 Baseline，定义新高度
需要特别说明的是，目前多任务榜单上仅有的两款模型（即 pi05_generalist 与 pi0_generalist），均为 RoboChallenge 团队训练的基准模型。这些模型在训练阶段被严格限制了数据规模（每个任务仅约 50 个样本），其主要目标是为社区提供一个可对标的基础参考分数。

因此，当前多任务榜单所呈现的成绩，更接近于在小样本多任务混合训练设定下的性能下限，而并非多任务具身智能模型所能达到的性能上限。我们诚挚邀请全球开发者在此基线之上展开探索与提升，共同推动多任务具身智能模型能力迈向新的高度。

3.3 不同维度的榜单

为了突破单一分数指标的局限性，更加立体、细粒度地评测模型能力，RoboChallenge 围绕 Table30 评测集构建了一套多维度的任务标签体系。

3.3.1 任务与标签的映射关系

Table30 中的每一个任务并非孤立存在，而是被赋予了三个维度的属性标签：机型、构型（Arm）与能力类型。

维度	标签	说明	任务数量	典型任务示例
机型（4种）	ALOHA	-	11	倒薯条 (pour_fries_into_plate)
	ARX5	-	11	放鞋上架 (place_shoes_on_rack)
	UR5	-	6	碎纸 (shred_scrap_paper)
	Franka	-	2	按按钮 (press_three_buttons)
构型（2 类）	Single-arm	单臂操作	19	放杯子 (put_cup_on_coaster)
	Two-arm	双臂操作	11	粘胶带 (stick_tape_to_box)
任务类型（9 个）	Precise3d	高精度 3D 操作	12	插网线 (plug_in_network_cable)
	Repeated	重复性任务	10	插花 (arrange_flowers)
	Bimanual	双臂/双手协同	8	扫垃圾 (sweep_the_rubbish)
	Manipulation	通用操控/抓取	6	开抽屉 (open_the_drawer)
	Classification	视觉分类	5	分拣电子产品 (sort_electronic_products)
	Multiview	多视角感知	5	找绿盒子 (search_green_boxes)
	Simple-pick	简单抓取	4	放鞋上架 (place_shoes_on_rack)
	Temporal	时序/长程记忆	3	扫二维码 (scan_QR_code)
	Softbody	软体/柔性物体	3	叠抹布 (fold_dishcloth)

3.3.2 按任务表现

参测 Top 9 模型在不同任务上的表现：

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

数据观察

- **Spirit-v1.5：** 在 9 个任务上获得最高成功率与最高过程分（含并列、SR > 0）。
  - 亮点任务：移动物体入盒（move_objects_into_box）成功率 SR 80% / 分数 Score 88.5。
  - 任务标签：repeated、single-arm、Franka

- **pi0.5：** 在 7 个任务上获得最高成功率与最高过程分（含并列、SR > 0）。
  - 亮点任务：堆碗（stack_bowls）成功率 SR 100% / Score 99.5，该任务也是当前平台提测次数最多的任务。
  - 任务标签：bimanual、two-arm、repeated、classification、ALOHA

- **wall-oss-v0.1：** 在 2 个特定任务上取得了 SR 100% / Score 100 的满分表现。
  - 亮点任务：按三个按钮（press_three_buttons）和堆叠色块（stack_color_blocks）。
  - 任务标签：
    - temporal、single-arm、Franka、repeated
    - simple-pick、single-arm、UR5、classification

3.3.3 按机型表现

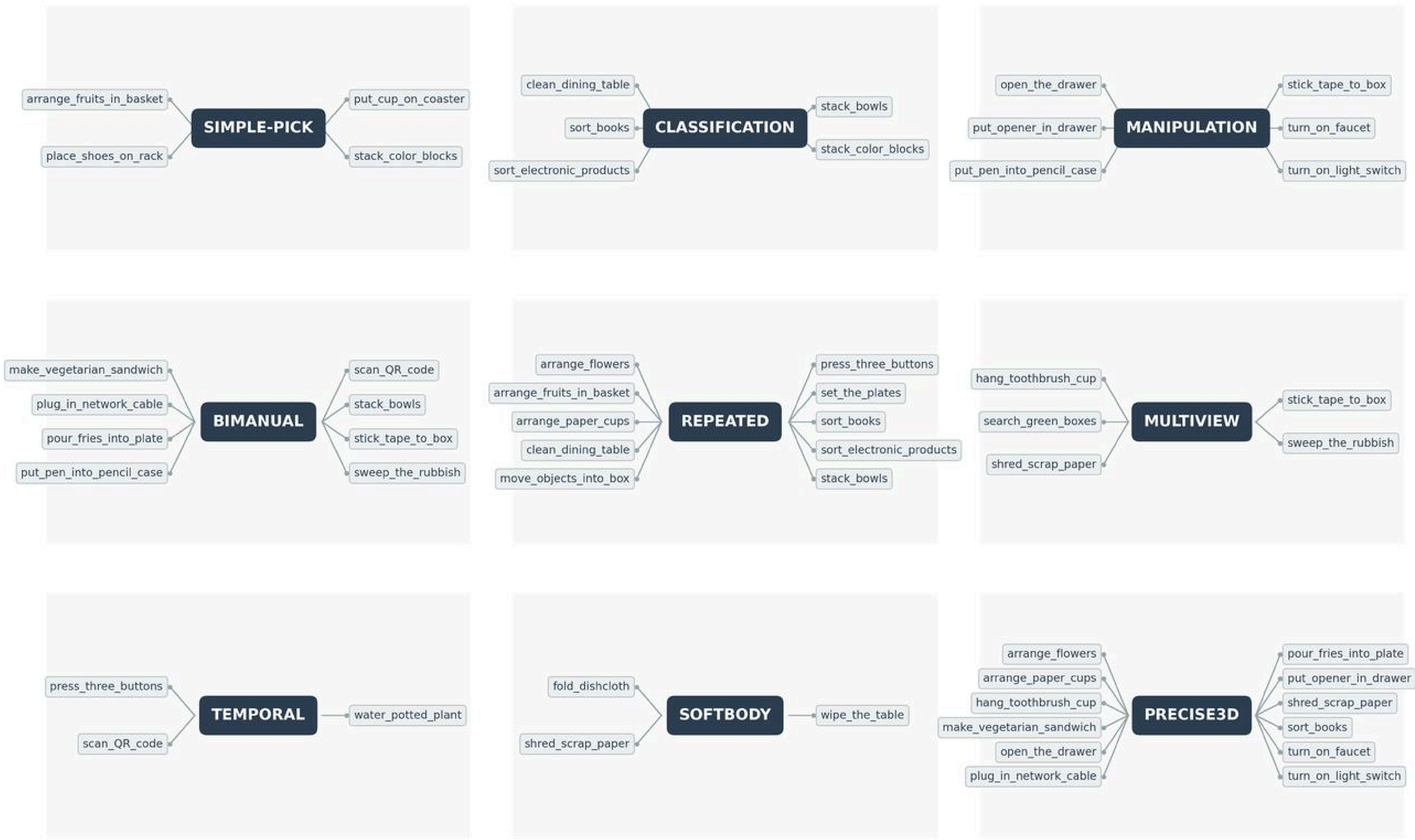
Table30 评测集包含 ALOHA、ARX5、UR5 和 Franka 四种主流机型，涵盖单臂与双臂构型。参测 Top 9 模型在不同机型上的表现：

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 按能力标签

通过基于任务的能力标签分析，我们可以看到本次上榜模型在不同能力维度上的具体表现（成功率）。

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


SIMPLE-PICK
arrange_fruits_in_basket
place_shoes_on_rack
put_cup_on_coaster
stack_color_blocks
CLASSIFICATION
clean_dining_table
sort_books
sort_electronic_products
stack_bowls
stack_color_blocks
MANIPULATION
open_the_drawer
put_opener_in_drawer
put_pen_into_pencil_case
stick_tape_to_box
turn_on_faucet
turn_on_light_switch
BIMANUAL
make_vegetarian_sandwich
plug_in_network_cable
pour_fries_into_plate
put_pen_into_pencil_case
scan_QR_code
stack_bowls
stick_tape_to_box
sweep_the_rubbish
REPEATED
arrange_flowers
arrange_fruits_in_basket
arrange_paper_cups
clean_dining_table
move_objects_into_box
press_three_buttons
set_the_plates
sort_books
sort_electronic_products
stack_bowls
MULTIVIEW
hang_toothbrush_cup
search_green_boxes
shred_scrap_paper
stick_tape_to_box
sweep_the_rubbish
TEMPORAL
press_three_buttons
scan_QR_code
water_potted_plant
SOFTBODY
fold_dishcloth
shred_scrap_paper
wipe_the_table
PRECISE3D
arrange_flowers
arrange_paper_cups
hang_toothbrush_cup
make_vegetarian_sandwich
open_the_drawer
plug_in_network_cable
pour_fries_into_plate
put_opener_in_drawer
shred_scrap_paper
sort_books
turn_on_faucet
turn_on_light_switch

能力标签和任务关系


Model Capabilities (SR%)
Spirit-v1.5
pi0.5
wall-oss-v0.1
pi0
pi05_generalist
RDT-1B
cogact
pi0_generalist
openvla-oft
manipulation
multiview
classification
bimanual
temporal
softbody
simple-pick
repeated
precise3d
20
50
80

数据观察

- **Simple-pick**（如 stack_color_blocks、put_cup_on_coaster 等任务）：这是目前上榜模型掌握得最好的领域，Spirit-v1.5 (85%) 、pi0.5 (80%) 和 wall-oss- v0.1 (77.5%) 均大幅超过 50%。
- **Manipulation**（如 stick_tape_to_box、open_the_drawer 等任务）：Spirit-v1.5 在此类任务上达到了 68.3% 的成功率，处于参测模型中的领先地位。
- **Classification**（如 sort_books、stack_color_blocks 等任务）：pi0.5 以 52% 的成功率反超 Spirit-v1.5 (48%)。
- **Softbody**（如 wipe_the_table、fold_dishcloth 等任务）：Table30 任务中的难题。即便是表现最好的 Spirit-v1.5，成功率也仅为 13.3%，其余模型普遍低于 10%。
- **Temporal**（如 press_three_buttons、scan_QR_code 等任务）：涉及长程记忆的任务模型表现普遍低迷，wall-oss-v0.1 在此类任务上取得 40% 的成功率。

四、 任务难度分析

RoboChallenge 榜单仅统计完成 Table30 评测集中全部 30 个任务的模型，为了更好地展示每个任务表现好的模型，我们从每个任务中筛选出表现突出的 Top 3 模型数据（取最优成绩）。

4.1 任务成功率分析

通过分析 Top 3 模型的成功率分布，我们将这 30 个标准化任务划分为三个不同的能力梯队。

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	rdt1	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 第一梯队：Hello World 级任务

定义：对于头部模型而言，此类任务已几乎没有难度，Top 3 模型的成功率均达到了 100%。这意味着目前的头部模型已基本掌握了此类技能。

代表任务：

- 堆碗（stack_bowls）：Top 1/2/3 成功率均为 100%，被测次数最多的“Hello World”级任务。
- 堆色块（stack_color_blocks）：Top 1/2/3 成功率均为 100%。

4.1.2 第二梯队：简单的任务

定义：此类任务对大多数头部模型较为友好。Top 1 成功率达到 90% 及以上，且 Top 3 的成功率也不低于 70%，这表明这些任务的门槛较低容易完成。

代表任务：

- 放鞋上架（place_shoes_on_rack）：Top 1（pi0.5）成功率 90%，Top 3 成功率 80%。
- 寻找绿盒子（search_green_boxes）：Top 1（Spirit-v1.5）成功率 90%，Top 3 成功率 70%。

4.1.3 第三梯队：特定模型的特长

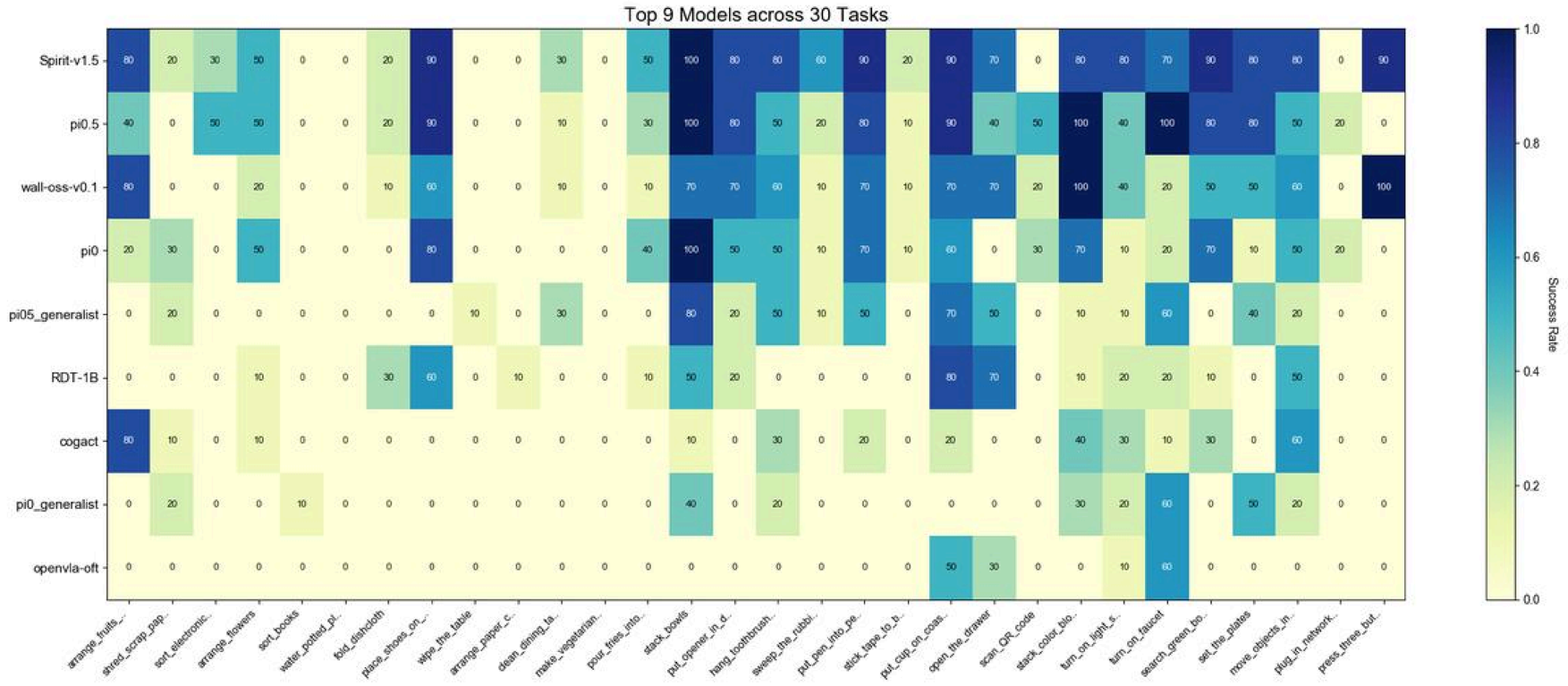
定义：此类任务呈现出极端的两极分化。Top 1 模型表现完美（100% 成功率），但后续名次的模型表现出现明显下跌。

代表任务：

- 按三个按钮 (press_three_buttons)：
  - Top 1（wall-oss-v0.1）成功率：100%
  - Top 3 成功率：降至 70%

4.2 失败任务分析

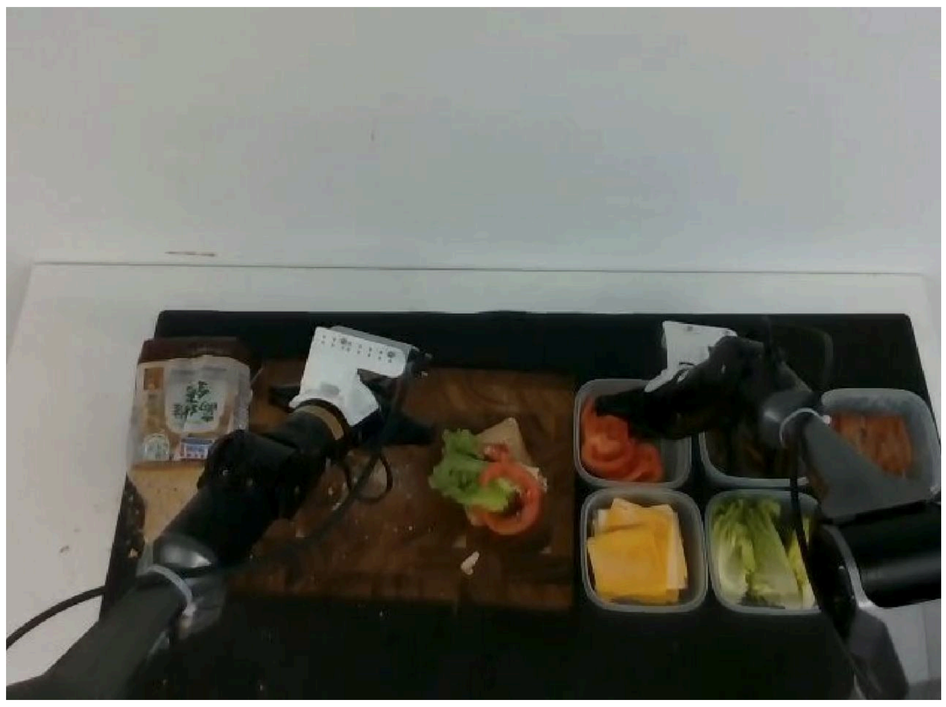
在上榜模型的评测结果（任务纬度）热力图中，有一些任务不仅平均成功率低，还多次出现所有参测模型全部为 0% 成功率的情况。例如，做素三明治（make_vegetarian_sandwich）、给盆栽浇水（water_potted_plant）等任务，我们形象地称其为“叹息之墙”。



下面是对五个失败任务的分析：

4.2.1 失败案例一：做素三明治（make_vegetarian_sandwich）

- **标签：** bimanual、two-arm、ALOHA、precise3d
- **记分点：**
  - 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.
- **数据：** 所有上榜的模型成功率均为 0%。
- **截图：**



- **复盘：** 该任务要求严格的顺序（放置面包 -> 蔬菜 -> 番茄 -> 面包）。评测显示，模型往往在第一步（左臂夹取物品）就出现数量错误或失败。
- **归因：** “一步错，步步错”。此类任务容错率极低，一旦初始步骤（如没抓起物品、拿错）失败，导致整个任务直接宣告失败。

4.2.2 失败案例二：给盆栽浇水（water_potted_plant）

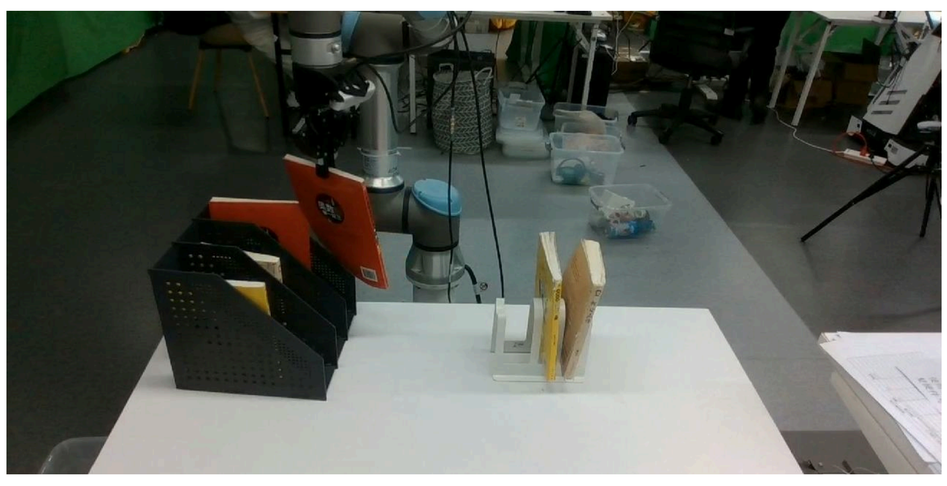
- **标签：** temporal、single-arm、ARX5
- **记分点：**
  - 4.0 points: Precisely grasp the handle of the kettle.
  - 2.0 points: Move the kettle to the position of the potted plant.
  - 3.0 points: Water the potted plant using the kettle.
  - 1.0 points: Place the kettle back to its original position.

- **数据：**所有上榜的模型成功率均为 0%。
- **截图：**

- **复盘：**模型能够成功完成前置动作（抓取水壶、移动至盆栽），但在重复浇水动作一段时间后，机械臂突然出现异常行为，伸直并大幅度移动，且未能执行将水壶放回原位的终止动作。
- **归因：**时序依赖缺失。长程任务要求模型维持对历史状态的记忆。一旦中间阶段出现状态丢失，模型就会陷入逻辑混乱，产生类似“幻觉”的随机动作。

4.2.3 失败案例三：整理书籍（sort_books）

- **标签：**precise3d、single-arm、UR5、repeated、classification
- **记分点：**
  - 1.5 * 3 points: Pick up one book.
  - 1.5 * 3 points: Put one book to the corresponding position.
  - 1.0 points: Reset the robotic arm.
- **数据：**上榜的模型最高成功率仅为 10%
- **截图：**



- **复盘：**模型在夹取第一本书失败后，直接跳过并尝试夹取第二本。但在执行时，机械臂意外夹取了第三本书，导致任务彻底混乱。
- **归因：**视觉感知精度不足与误差累积。此类任务要求极高的视觉分割能力（区分紧挨着的书本）。一旦视觉判断出现微小偏差，或者前一步操作失败，误差会在多步操作中被放大，最终拉低整体成功率。

4.2.4 失败案例四：叠抹布（fold_dishcloth）

- **标签：**softbody、single-arm、ARX5
- **记分点：**
  - 4.0 points: Fold the dishcloth first time.
  - 4.0 points: Fold the dishcloth second time.
  - 2.0 points: Place the folded discloth on the front-left position.

- **数据：** 上榜的模型最高成功率为 30%
- **截图：**

- **复盘：** 机械臂在尝试对折抹布时，由于布料受力后形状发生不可预测的变化，导致机械臂多次尝试均未能完成对折，最常见的情况是变成了拖着抹布走。
- **归因：** 物理形变预测难。相比于刚体，软体的操作难度呈指数级上升。当前算法难以精确把控对柔性物体的操作力度与抓取位置，是行业公认的难点。

4.2.5 失败案例五：排列纸杯（arrange_paper_cups）

- **标签：** repeated、single-arm、ARX5、precise3d
- **记分点：**
  - 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.
- **数据：** 全场最佳模型成功率为 20%。
- **截图：**

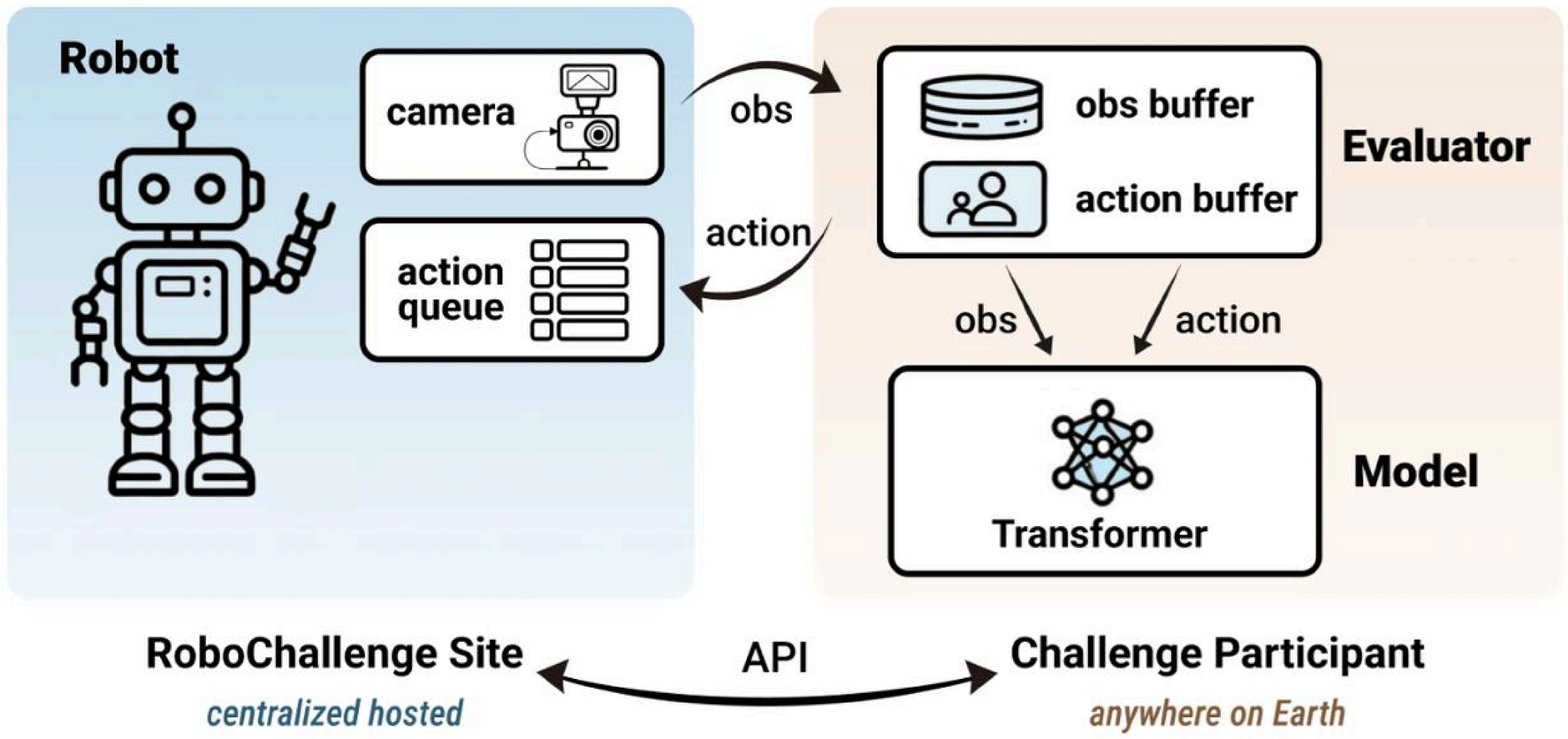
- **复盘：** 模型其实已经成功完成了前 80% 的工作（成功夹起并套好了 4 个纸杯）。但在最后一步试图将叠好的杯子放入架子时，机械爪意外将杯塔推倒。随后，机械臂在桌面上多次尝试重新抓取倒下的杯子，但均告失败。
- **归因：** 复杂的堆叠任务要求全流程的完美控制，任何一个动作（如释放夹爪）的微小抖动，都可能摧毁之前所有的努力。

五、RoboChallenge 评测体系与标准化流程

在具身智能领域，真机评测的标准化与公平性长期以来一直是制约技术横向对比的关键瓶颈。RoboChallenge 作为原力灵机 Dexmal 与 Hugging Face 联合推出的全球首个具身智能大规模真机评测平台，构建了一套简单易用的远程真机评测体系。接下来，我们将详细介绍其架构设计与评分标准。

5.1 基于 API 的远程真机评测平台

不同于需要提交模型权重或 Docker 镜像的评测模式，RoboChallenge 平台采用了用户端推理的架构设计。在这一模式下，模型运行在参测方本地的计算资源上，并通过标准化在线 API 与评测平台的本体进行交互。

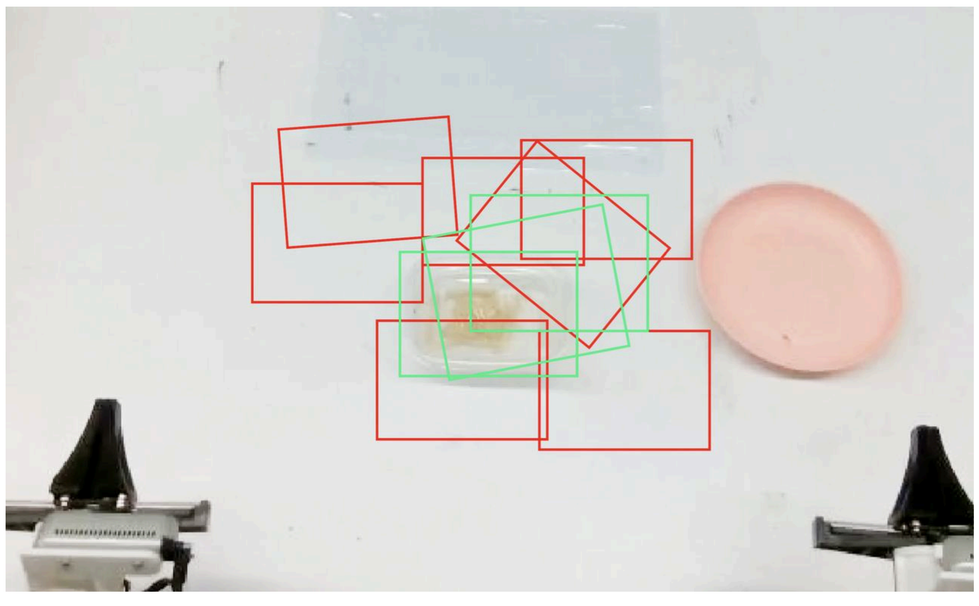


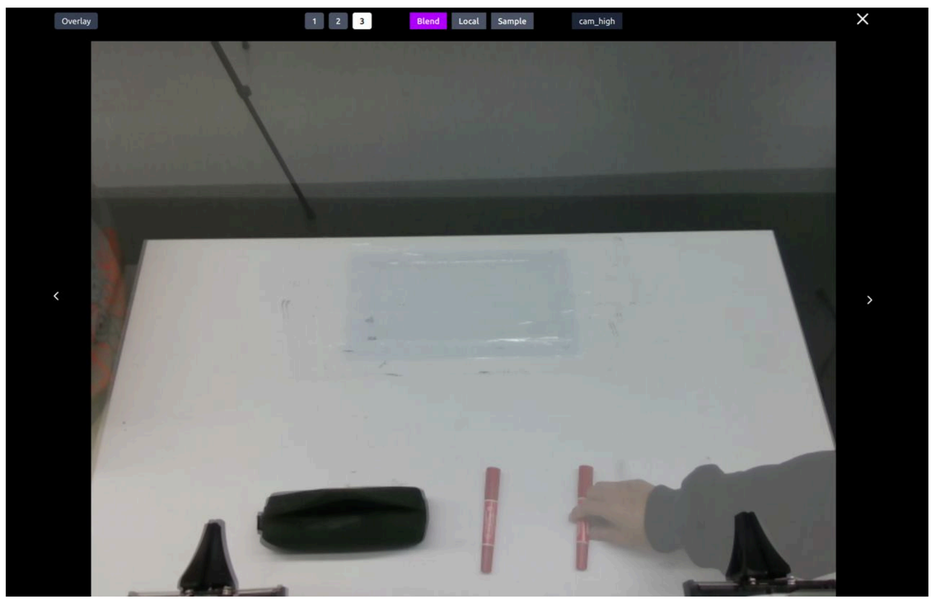
用户无需处理复杂的环境配置迁移问题，只需通过 API 接收带有精确时间戳的观测数据，包括 RGB 图像、深度图和本体感知（proprioception）数据，以及动作指令。

5.2 消除人为偏差：视觉输入匹配方法

在真机评测中，最大的变量往往来自操作员对测试场景的布置。研究发现，有经验的操作员可能会在无意识间将物体放置在模型更容易成功的“甜点区域”，从而造成评测结果的偏差与失真。

为了消除这一系统性偏差，RoboChallenge 引入了视觉输入匹配（Visual Task Reproduction）方法。在每次测试重置阶段，系统会从训练数据中提取一帧参考图像，并将其半透明地叠加在操作员的实时预览画面上。测试人员需不断调整物体位置，直至实时画面在视觉上与参考图像高度重合。借助这种方式，每次测试的初始状态分布都能与训练数据保持一致，从而显著提升评测的稳定性与公平性。





5.3 多维度量化评分标准

为了更精细地衡量模型性能，RoboChallenge 采用了成功率（Success Rate）过程分（Progress Score）的双重评价体系。

- **成功率（Success Rate）：** 衡量整体任务是否完成的指标

- **过程分（Progress Score）：** 衡量在任务执行过程中向成功目标推进的程度。RoboChallenge 将每个任务拆解为若干执行阶段，每完成一个阶段即可获得对应分数。即便最终任务失败，模型在中间阶段取得的部分进展也会被量化记录。

- **重试惩罚：** 针对实际操作中常见的不稳定性，系统引入了重试扣分机制。如果本体在某一阶段尝试失败（如抓取脱手）并发起重试，每次将扣除 0.5 分。当某一阶段得分被扣至负数，或连续失败次数超过 4 次，该轮测试将自动终止。

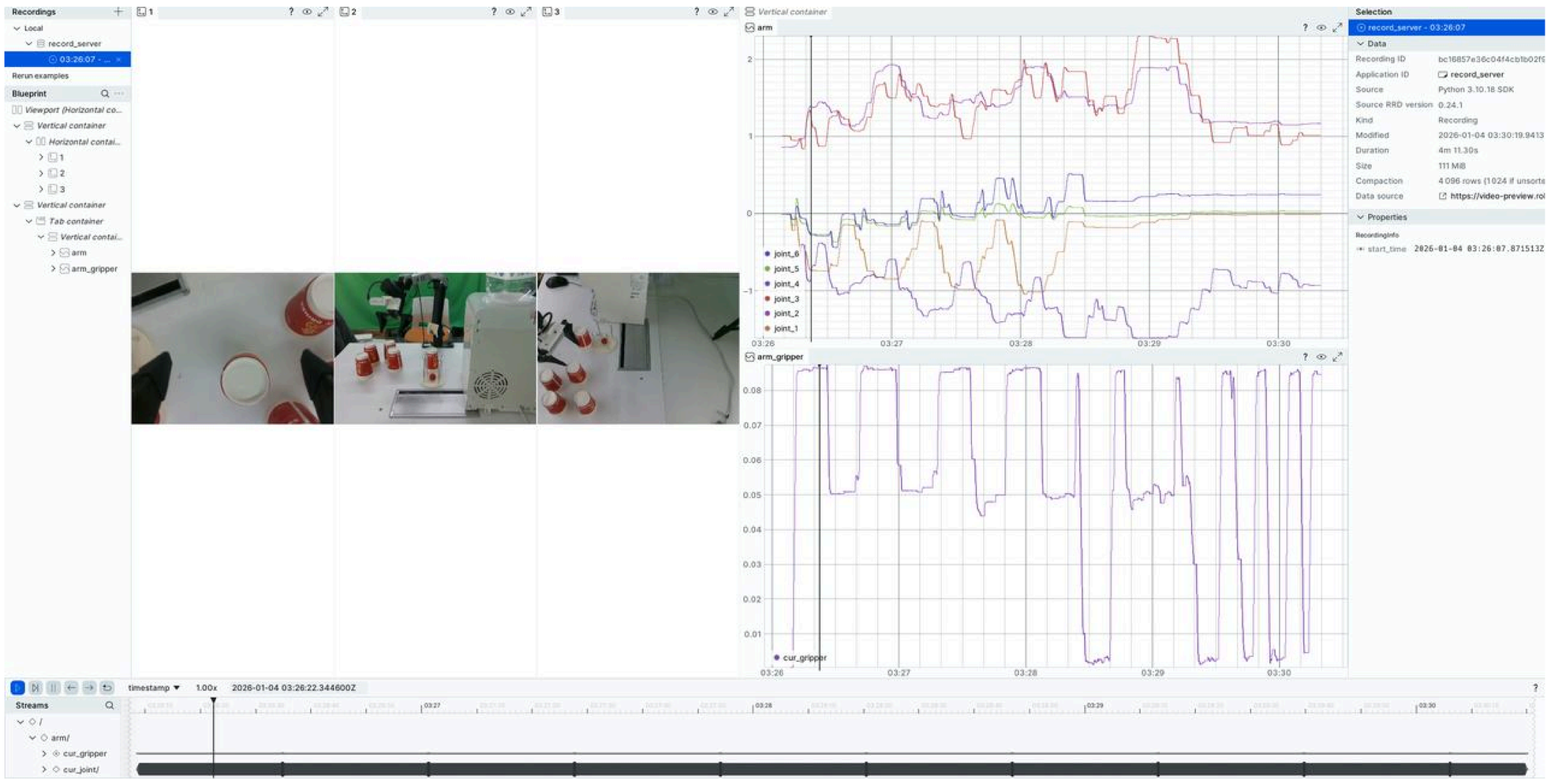
5.4 开源真机数据集

RoboChallenge 平台为 Table30 基准中的每一个任务提供了丰富的真机演示数据，单个任务包含多达 1000 条完整轨迹（Episodes）。

所有数据集均托管于 Hugging Face，对社区完全开放。数据不仅包含原始的视频文件和 JSON 格式本体状态信息，还提供了转换脚本，支持将数据转换为 LeRobot 等主流数据格式，极大降低了开发者的使用门槛。

5.5 提交评测与透明度

在提交机制上，平台区分了特定任务（Task-specific）与多任务（Multitask）模型两种设定。同时，为了促进社区交流，RoboChallenge 会在官网公开参与者的评测录像与机器日志（RRD 格式），研究者可以相互查阅、复盘并分析失败案例，从而推动算法的进步。



六、RoboChallenge 合作伙伴生态

6.1 社区发展的历程

原力灵机 Dexmal 与 Hugging Face 共同发起 RoboChallenge 之后，迅速在行业内引发强烈共鸣并吸引广泛参与。智源研究院、智元机器人、Qwen、星海图、自变量、清华大学、西安交通大学、GOSIM 国际国内合作伙伴进一步携手，共同推动生态建设，并于 2025 年 11 月 20 日正式成立 RoboChallenge 组委会。这标志着具身智能真机评测的开源协作不仅迈入了标准化的新阶段，更以“开放共同体”的行业共创模式，为具身智能技术的落地与迭代注入新动能。

组委会的成立旨在将 RoboChallenge 升级为行业级公共基础设施，通过标准化、常态化的运营机制，推动具身智能评测从“分散实验”迈向“共识共建”。未来，组委会将携手产业界、学术界与开源社区，构建透明、高效、可信的具身智能评测生态，助力行业评测标准的加速形成与落地。

6.2 组委会启航以及组织架构

成立的 RoboChallenge 组委会，将采用指导委员会（Steering Committee） + 工作组（Working Group）的双层架构，确保高效执行与透明治理；组委会作为决策中枢，遵循“开放、协作、共识”原则，统筹全局；下设四大核心工作组，作为具体落地的“执行引擎”。

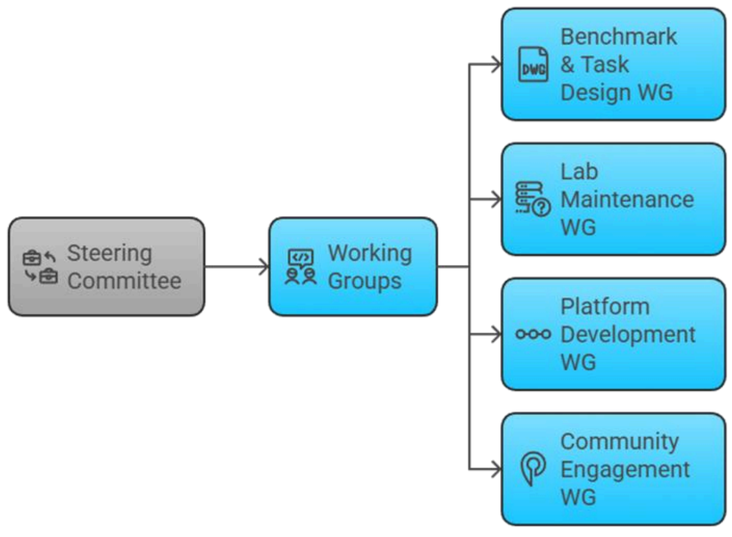
- 基准与任务设计工作组（Benchmark & Task Design WG）：聚焦评测体系的“核心大脑”。该工作组负责设计与优化具身智能基准、典型任务及配套数据集。

- 实验室维护工作组（Lab Maintenance WG）：扮演“硬件管家”角色。该组负责接收合作伙伴捐赠的测试机器人，维护实验室环境，并执行标准化测试。通过统一硬件条件、测试流程，确保不同机器人的评测结果可对比、可复现。

- 平台开发工作组（Platform Development WG）：打造“数字底座”。该组负责开发与维护 robochallenge.ai 网站，开发 API 服务与数据分发工具，让全球开发者能便捷上传测试数据、获取评测结果，形成“测试-反馈-优化”的闭环生态。

- 社区共建工作组（Community Engagement WG）：承担“放大器”功能。通过举办黑客马拉松、学术研讨会、线上线下活动，吸引更多开发者、企业、高校加入，推动技术交流与需求对接。

RoboChallenge Organizational Structure



申请加入 RoboChallenge 工作组社区的组织意愿持续高涨。截至 2026 年 1 月 23 日，我们已正式收到来自多家机构的加入申请，包括：四川省具身智能机器人训练场、中移杭研、元点智能、原力无限机器人、智平方、光轮智能、深圳星际光年、北京人形机器人创新中心、千寻智能、蚂蚁灵波科技、极佳视界等。

2026 年 1 月，星海图捐赠了 R1 Lite 到 RoboChallenge 组委会的真机实验室，进一步丰富了 RoboChallenge 的硬件构型库，为后续开展跨本体，多场景的真机评测提供了有力的硬件支持。

随着社区规模的不断扩大，RoboChallenge 工作组社区正逐步形成一个高频互动、实质共建的协作网络。社区成员围绕评测任务设计、指标体系、真机复现等核心议题持续展开深入讨论与贡献，共同推动具身智能评测体系的完善与行业整体进步。

6.3 治理原则： 以共识驱动可持续未来

除了具体的执行架构，RoboChallenge 组委会的运作原则也值得关注；其核心承诺包括三点：测试开放可复现、社区包容非竞争、贡献开源可追溯。这意味着，任何参与方的技术细节、测试数据、改进方案都需在框架下公开，避免“闭门造车”；具身智能的评测不是淘汰赛，而是共建场。

通过开放治理，RoboChallenge 组委会希望吸引更多研究机构、企业、初创团队加入——它们往往更贴近真实需求，能为评测体系注入鲜活场景。例如，家庭服务机器人企业可能关注“多语言指令响应”的评估维度，这些都将被工作组吸纳，最终反哺整个行业。

6.4 社区对未来 RoboChallenge 发展的反馈和建议

在过去 3 个月阶段中，RoboChallenge 与研究者、开发者、企业合作伙伴持续互动，积累了大量对未来发展的建设性建议。综合 RoboChallenge 组委会， 四个工作组以及广大社区伙伴们的这些反馈，我们归纳出以下关键方向与具体建议点：

1）提升 Zero-Shot 能力评估的覆盖与体系化

RoboChallenge 社区多次提到，当前真实机器人评测往往集中在预训练或特定微调后的任务上，但对于“未见任务 / 零样本执行能力（Zero-Shot）”的评估体系仍不完善。

- 制定 Zero-Shot 能力标准与任务分层指标；
- 开发一套专门针对指令理解 + 环境感知 + 动作生成的 Zero-Shot benchmark；
- 对比不同模型在 Zero-Shot 下的表现差异，并公开可复现评测脚本与分析报告。

2）构建真正意义上的通用多任务（Multitask）评测体系

社区普遍希望 RoboChallenge 能支持多任务通用模型的系统性评测，不仅是单任务排行，而是整体多任务协调能力。

- 设计跨任务组合方案，例如按任务类型（拾取 / 装配 / 推理 / 复杂序列动作等）划分标签；将多个原子任务组合成为一个复杂的组合型任务；
- 推出“通用策略挑战赛”，鼓励提交可在多任务上同时保持性能的模型；
- 设定统一评分机制包括准确率、连续性、执行效率等多维评估指标。

3）真实环境 vs 仿真评测效果对比曲线可视化

合作伙伴反馈中一项重要诉求是：同一模型在仿真器与真实环境中表现的落差，这对于理解模型泛化能力至关重要。

- 1:1 对应增加仿真和真机的测试任务集。如目前已经有 Table30，就可以考虑生成 Table30 的仿真环境；
- 如有一套仿真任务，就可以对应支持真机的任务集；
- 定期发布对比曲线图，如 “仿真得分 vs RoboChallenge 真实环境得分”，用统一任务集与统一得分标准可视化模型在两者之间的差距；
- 提供基础可视化模板与定制化报告功能，便于社区自行生成对比曲线；
- 持续积累此类横向对比数据，为未来研究提供真实迁移难度量化基础。

4）加大大赛协办与产业合作支持力度

社区和机构反馈，希望 RoboChallenge 能成为具身 AI 领域更具影响力的大赛平台，比如与顶级会议（ICRA / RSS / NeurIPS Robotics Track）或标志性赛事合作。

- 主动邀请高校、研究院、产业联盟成为大赛协办单位；
- 与仿真平台及机器人本体供应商合作，提供官方 benchmark 奖项与实机测试支持；
- 探索将 RoboChallenge 评测纳入一些主流国际机器人赛事的官方评分子赛道。

5）持续扩展 Benchmark 规模与多样性

RoboChallenge 现有的 Table30 是良好开端，但社区呼声聚焦于更多维度、更复杂场景、更具挑战性的任务集合。

- 引入更高层次任务 benchmark，例如复杂装配、动态场景动作、多人协作任务等；

- 与社区共同构建“Bench30+ / Bench100 系列任务集”，包括由社区提交并经过审查的任务；
- 定期发布 benchmark 库更新与 leaderboards，强化数据集生态活跃度。

6）增加更多机器人机型与适配接口

合作伙伴反馈指出，真实机器人种类繁多，而适配不同机型往往需要重复劳动。

- 支持更多的本体，灵巧手等机器人设备接入 RoboChallenge 评测平台；
- 优先纳入社区需求高、工业应用广的机型；
- 发布“机型适配共建计划”，鼓励硬件厂商提交驱动与适配代码，并纳入官方支持。

7） 增加评测场景维度与现实复杂性

现实世界的具身任务常伴随不确定性、动态变化、狭小空间等挑战。社区建议 benchmark 不应局限于静态桌面集。

- 发布多场景 benchmark，例如厨房、家居、仓储、零部件装配线等场景；
- 引入部分动态元素（可移动障碍、变化的目标位置等）。

8）增加排队机制的透明性

从用户视角来看，用户不知道需要等多久才能评测上，以及很多人提交测试的时候，无法保证每一位用户都能测上。建议如下：

- 透明化队列系统：设置一个实时、透明的排队列表（or waitlist），或在后台显示当前待测任务数。按照提交顺序自动触发测试，减少人工沟通成本，使评测流程更符合国际开源社区的协作习惯；
- 提交额度管理：为平衡服务器负载与资源公平性，建议限制单个账号每日的提交次数。这既能有效控制总任务量，也能引导参赛团队在本地充分验证后再提交，提升整体评测效率。

9）构建严苛的泛化性与“盲测”体系

从评测的科学性与严谨性来看，为了防止模型对特定环境“过拟合”或通过“背题”获得高分，可以建立多维度的动态考察机制。建议如下：

- 全方位的泛化性考察： 建立从物体、位置到场景的三级泛化测试标准。通过随机变换目标物体的摆放位置、引入未见过的同类物体以及改变背景光照或纹理，全面检验模型是真正理解了“语义”，还是仅仅记住了“像素”；
- 严格保密硬件配置： 实施评测环境参数的“盲测”机制，不对外公开摄像头的俯仰角、安装高度等关键数据。迫使模型必须依赖视觉感知与适应能力来完成任务，而非利用已知的参数先验进行硬编码作弊，确保评测结果的真实与公正。

七、平台运营复盘

RoboChallenge 作为一个开创性的真机评测平台，在上线后经历了从“原型验证”到“规模化服务”的跨越式发展。面对真实物理世界的复杂性与激增的行业需求，我们直面挑战，通过持续的复盘与迭代，不断夯实评测体系的科学性与权威性。

7.1 产能瓶颈问题

平台上线初期，系统设计的日均承载阈值为 800 次真机测试（Rollouts）。然而，实际评测需求的增长速度远超预期，特别是在 12 月份，日均真机测试（Rollouts）需求常态化维持在 1000 至 1500 次区间，呈井喷之势。面对严峻的供需缺口，运营团队通过优化调度算法与扩展服务时段等多维手段，成功将系统实测承载量提升了 50%。

7.2 测试一致性难题

尽管 RoboChallenge 在设计之初便确立了包含任务设计、数据标准及打分规则在内的标准化体系，但真实物理环境的“非受控性”依然是最大的变量。但我们必须承认：在非受控的物理世界中，绝对的一致性是难以企及的理想状态。 实测数据显示，即便在严格控制的实验室环境下，微小的光照波动、操作员着装颜色的差异，都可能导致同一模型在两次提交中产生成绩波动。目前，我们尚未能完全消除这些物理“噪声”，但这些实测问题案例不仅帮助我们更好的认识到了模型对各种干扰因素的敏感度，也对未来评测集设计的科学性和评测的严谨性积累了宝贵经验。

7.3 评分二次审核

针对人工评测可能存在的主观偏差，我们意识到单一的打分表已无法满足严谨性要求。为此，平台构建了包含大量标准案例的评测知识库，并引入了二次审核机制与申诉仲裁渠道。虽然我们无法保证 100% 的初判准确率，但通过这套纠错流程，我们显著降低了结果的统计偏差，确立了有错必纠、有据可查的评分原则，帮助 RoboChallenge 评测结果更加严谨。

7.4 信任机制与开源呼吁

鉴于 RoboChallenge 平台采用的用户端推理的架构特性，平台作为第三方，客观上无法从技术层面验证用户私有模型架构的真实性，也无法完全规避针对性调优的问题。这是远程真机评测面临的局限。 因此，RoboChallenge 诚挚呼吁社区拥抱开源与透明。我们始终坚持中立的服务原则，但唯有通过代码与权重的公开，结合真机评测的实证，才能彻底打破猜疑。

7.5 远程调用影响产能上限

RoboChallenge 采用用户端推理架构，虽然极大降低了用户的环境迁移成本并保障了模型资产安全，但也带来了不可忽视的效率折损。在实际运行中，真机评测的整体效率不再单方面取决于机器人执行速度，而是深度耦合了网络传输延迟与用户端模型推理耗时。即便是毫秒级的网络波动或用户端的算力不足，都会直接转化为机器人的“空转等待”，从而拉长单次任务的占用时间，进而制约了平台整体的日均吞吐量。 如何进一步压缩通信时延、提升异地推理的协同效率，将是我们下一阶段通过技术攻坚解决的核心课题。

八、评测者实战经验和洞察

8.1 评测方实战复盘 - 来自社区志愿者的实测经验

社区志愿者来自 TAO-DualArmVLA 模型团队的成员分享了其在 RoboChallenge 评测中的实战经验。

1. 数据工程：高质量的“燃料”是成功的一半。在模型训练之前，我在原始数据治理上投入了大量精力，实测证明这一投入有效提升了真机评测的分数。
   - 原始数据深度观察：通过可视化手段对视频及传感器数据进行直观分析，精准掌握了数据的分布规律、尺度范围及异常值边界，为后续处理提供了依据。
   - 信号清洗与离散化：针对原始数据中夹爪传感器采集的连续波形信号（通常包含不必要的噪声与干扰），为了提升训练效果，我将其转换为代表“开/合”两种状态的方波信号。这一处理显著增强了模型对末端执行器状态判断的稳定性。

2. 策略选择：由点及面，寻找突破口。面对多任务并发的评分体系，盲目追求“全面开花”往往是低效的。我采取了由点及面的攻坚策略：
   - 聚焦高分任务（关键路径法）：优先锁定评分权重高且易于突破的任务建立基线，集中资源打通关键路径。
   - 交叉比对与策略评估：在单点突破的基础上，进行策略的交叉比对与评估，快速筛选出具备迁移价值的通用模块。

3. 模型训练：科学实验，迭代优化。我将模型训练视为一个严谨的科学实验过程，摒弃了随机试错的模式：
   - 系统化实验档案：使用工具详细记录每一次实验的超参数、代码版本、数据集版本及对应的性能指标，实现全链路可追溯。
   - 严格控制变量：坚持采用控制变量法进行调参，每次仅改变 1-2 个关键参数。这种方式能清晰归因性能变化的来源，并最终形成清晰的实验报告，确保持续优化的方向正确。

8.2 完成评测方 - 来自千寻智能团队的评测经验

在 Spirit-v1.5 的研发与评测过程中，我们针对模型在复杂物理环境下的泛化瓶颈进行了系统性分析，并在算法架构与控制策略上实现了以下关键演进：

- 针对精细操作的解耦策略：在评测中我们观察到，模型在执行高精度任务时容易产生过拟合的问题，即过度依赖机械臂的反馈状态（State）而忽视了实时的视觉引导，导致末端位姿在关键时刻出现偏差。为此，我们创新性地提出了 Mask Z 策略：在输入状态中屏蔽掉高度（Z 轴）信息，强制模型通过视觉来实时推断物体高度。这一改动显著增强了模型在不同桌面环境下的感知泛化性，有效降低了定位误差。
- 极限位置下的动力学稳定性优化：针对 Aloha 等硬件平台在处于工作空间极限位置时容易出现的末端抖动问题，我们在推理侧引入了动作平滑机制。通过对输出的 Action Chunk 进行时序平滑处理，保证了机械臂在极端位置运行的稳定性，大幅提升了任务完成的连贯率。
- 视觉分辨率对空间感的影响：打榜结果暴露出模型在面对微小物体抓取时，依然存在空间感不足导致的空夹现象，这主要受限于预训练阶段视觉编码器的输入分辨率。我们未来计划引入更

高分辨率的视觉编码器来解决这一问题。

感谢官方搭建了高水平的具身智能评测平台，让我们能在一个公正、公开的环境下验证 Spirit-v1.5 的通用性。特别感谢在模型部署与真机调试过程中提供技术支持的官方团队，你们的反馈是驱动我们模型迭代的重要动力。展望未来，我们期待与 RoboChallenge 共同推动和见证具身智能的 GPT-3 时刻。

8.3 完成评测方 - 来自自变量的评测经验

在完成 RoboChallenge Table30 全流程评测的过程中，Wall-oss 团队不仅验证了模型在多维任务中的边界能力，更在与真实物理世界的交互中积累了宝贵的工程经验。以下是我们在实操过程中的反馈：

- 针对推理延时的全链路优化策略： 在初期评测中，我们受限于远程通信与推理的累积时延，导致任务执行效率低下。为此，我们实施了端到端时效性对齐策略：一方面向 RoboChallenge 平台同步精确的单步时间预算以减少同步等待；另一方面通过适配图像传输分辨率与深度优化模型推理算子，成功将全流程执行速度提升了 5 倍以上。这一工程化调优不仅解决了超时问题，更为后续大规模自动化测试奠定了基础。

- 非受控环境下的 Sim-to-Real 对齐挑战： 真机评测展现了物理世界的高熵特性。我们观察到，背景环境的微小不一致或物体初始状态的随机改变，都会引发模型性能的非线性波动。这种环境噪声对模型鲁棒性的影响边界，是当前学术界与工业界亟需共同量化的课题。我们建议未来引入更细粒度的环境干扰测试，以精准测量模型的抗扰动能力。

- 评分体系的细粒度量化建议： 针对评测中偶尔出现的评分歧义，我们建议在现有的打分标准基础上，引入明确的加分项与扣分项细则。通过将评分维度从“结果导向”下沉至“动作细节”，不仅能进一步降低人为判罚的主观差异，也能帮助开发者在离线测试中更精准地与线上标准对齐。

RoboChallenge 的 Table30 任务集覆盖了精细视觉感知、时序逻辑理解及多本体适配等具身智能的核心能力，感谢主办方团队对构建标准化评测体系的努力。过程中我们和主办方也一起解决了机器人跨本体泛化中的末端控制问题，未来我们希望与 RoboChallenge 携手继续推动评测标准的进化，期待共同见证具身智能迈向通用化。

8.4 进行中的评测方 - 来自极佳视界 GigaBrain 团队的评测经验

极佳视界在基于完全自研和全栈开源的具身大模型 GigaBrain 在 RoboChallenge Table30 任务中的首轮摸底测试中，从实战角度总结了以下关键经验，旨在为社区提供具身模型训练与推理优化的参考：

- 基座模型预训练覆盖度决定能力边界：在评测初期，我们发现任务的难易程度与模型表现并不总是正相关：部分直觉上的“简单任务”因训练数据分布差异，成功率可能仅为 80%；而部分“复杂任务”若命中了基座模型的预训练域（如空间表征或特定操作语义），反而能表现出极佳的鲁棒性。这一现象表明，基座模型的预训练数据广度直接决定了下游任务的泛化门槛，仅靠后训练难以完全弥补基座在某些长尾分布上的认知缺失。极佳视界正在践行和引领的世界模型技术路径，可以从根本上解决基座模型的泛化性问题，不仅在 RoboChallenge 的任务上得以体现，也使得真正实现物理 AGI 成为可能。

- 建立“开环指标-闭环实测”的映射关系：鉴于真机评测资源的稀缺性与时间成本，直接依赖全量真机测试效率较低。我们团队建议优先进行充分的开环评测，通过分析 Loss 曲线、开环动作轨迹与最终真机成功率之间的交叉关联，总结出一套可靠的预测规律。利用这一规律指导模型筛选（Checkpoint），可显著减少无效的真机提交，实现评测效率倍增。

- 推理参数需针对“远程真机”进行精细化适配：不存在“一套参数通吃”的通用解。针对 RoboChallenge 的跨本体（不同机型）与跨任务特性，必须对动作执行长度、控制频率、末端策略等推理参数进行专项调优。此外，由于 RoboChallenge 采用远程评测模式，网络延迟与波动是不可忽视的干扰因素。

- 用好时序信息可以起到事半功倍的效果：在多步骤长程任务中，常出现不同阶段视觉输入极其相似的“状态混淆”现象（如物体拿起前后的静止状态）。单纯依赖当前帧往往导致决策死循环。为了解决这个问题，我们创新性的引入强时序状态转换模块，它能帮助模型利用历史上下文信息辨析当前状态，显著提升任务执行的连贯性与流畅度。

RoboChallenge 是全球首个具身智能大规模真机评测平台，具有很强的落地指导意义。接下来希望能和 RoboChallenge 一起携手共创，共同推动物理智能的基准提升，共同见证物理 AGI 时刻的到来。

致谢：

原力灵机 Dexmal 团队为 RoboChallenge 平台提供了真机集群资源和长期的运营人力保障，使得大规模、标准化的真机评测成为可能。

RoboChallenge 组委会汇聚了产学研各界的核心力量，围绕评测指标体系设计、评测数据的分析与复盘、平台未来演进路径以及行业发展趋势等关键议题，积极参与讨论并持续贡献洞见，共同推动具身智能评测生态的成熟与演进。

编写委员会

陈阳、孙雪、范浩强、马腾、黄小天、张恩文、刘凯、郭俊良、张蒲石、王鹏伟、王铁震、江波、何晓萌、陈乙宽、王洋、朱政、郁葱葱