摘 要

手内旋转要求灵巧手在不重新抓取物体的情况下持续改变其姿态，是精细操作中的一项基础能力。面向低成本 LEAP Hand 在无视觉、无触觉条件下的连续旋转部署，本文在 Isaac Lab 中建立圆柱体连续手内旋转任务、奖励函数和统一评测协议，并将 HORA/RMA 两阶段适应思路改造为适用于 LEAP 平台的 8 维技能先验表示。为缓解第二阶段部署策略在分布外圆柱参数下的退化，本文在训练中扩大质心偏移覆盖范围，并引入教师——部署动作一致性损失，最终得到仅依赖本体感觉历史的部署策略。

仿真中，第一阶段教师策略能够在 20 s 回合内维持稳定的圆柱体连续旋转。增强后的 Deploy-Refined 在分布内评测中成功率为 1.0000，在常规分布外评测中将成功率由 0.8750 提高到 0.9141，存活时间、净旋转圈数和轴向角速度也同步改善。消融实验显示，动作一致性约束贡献了第二阶段补强的主要增益，放宽质心偏移分布则增加了高风险样本覆盖。实机测试中，同一策略可以在真实 LEAP Hand 上闭环运行，并在标准圆柱体、尺寸变化圆柱体和细高饮料瓶上产生可重复的低速受控旋转。本文结论限于圆柱体参数泛化和少量近圆柱对象迁移，任意多物体稳定旋转仍是后续研究目标。

关键词：LEAP Hand；手内旋转；强化学习；本体感觉自适应；Sim2Real

ABSTRACT

In-hand rotation requires the dexterous hand to continuously change its posture without re-grasping the object, which is a basic ability in fine operation. For the continuous rotation deployment of low-cost LEAP Hand under non-visual and non-tactile conditions, this paper establishes a cylinder continuous hand rotation task, reward function and unified evaluation protocol in Isaac Lab, and transforms the HORA / RMA two-stage adaptation idea into a 8 dimensional skill prior representation suitable for LEAP platform. In order to alleviate the degradation of the second-stage deployment strategy under the distributed outer cylinder parameters, this paper expands the coverage of centroid offset in training, and introduces the loss of teacher-deployment action consistency, and finally obtains the deployment strategy that only depends on the history of proprioception.

In the simulation, the first-stage teacher strategy can maintain a stable continuous rotation of the cylinder in the 20 s round. The enhanced Deploy-Refined has a success rate of 1.0000 in the intra-distribution evaluation, and the success rate is increased from 0.8750 to 0.9141 in the conventional out-of-distribution evaluation. The survival time, the number of net rotations and the axial angular velocity are also improved simultaneously. Ablation experiments show that the motion consistency constraint contributes to the main gain of the second stage reinforcement, and the relaxation of the centroid offset distribution increases the coverage of high-risk samples. In the real machine test, the same strategy can run in a closed loop on the real LEAP Hand, and generate repeatable low-speed controlled rotation on standard cylinders, size-changing cylinders, and slender beverage bottles. The conclusion of this paper is limited to the generalization of cylinder parameters and the migration of a small number of near-cylinder objects. The stable rotation of any number of objects is still the goal of subsequent research.

Key words: LEAP Hand; in-hand rotation; reinforcement learning; proprioceptive adaptation; sim-to-real

目 录

1 绪论

1.1 研究背景与意义

随着机械手系统逐步进入家庭服务、柔性装配和开放环境操作场景，手部操作研究的关注点也在变化：以前更多关心“能不能抓住”，现在则更看重“抓住以后还能不能继续处理”。相比平行夹爪，多指灵巧手具有更丰富的运动自由度和接触构型，更适合完成姿态调整、工具使用和精细装配等任务[1–4]。在这样的背景下，手内操作逐渐成了灵巧操作研究中绕不开的问题之一。

手内旋转是手内操作里最典型、也最能体现连续接触控制难度的基础技能之一。它要求机器手在不重新抓取物体的前提下，仅靠手指之间的协调接触持续调整物体姿态。和一次性的抓取或搬运不同，这个过程里同时包含了高维控制、接触切换、摩擦不确定性和姿态稳定性等问题，因此对策略稳定性、感知质量和控制精度都提出了更高要求。对实际应用来说，手内旋转可以直接服务于工件对位、工具使用和姿态调整等场景，工程价值比较明确。

传统解析模型在这个问题上并不轻松。手内旋转过程中常常伴随复杂的非线性接触、摩擦不确定性以及频繁的接触模式切换，精确建模的代价很高。只要对象质量分布、摩擦特性或外部扰动发生变化，基于固定模型设计的控制器就可能需要重新标定。阻抗控制等经典方法确实为接触稳定性分析提供了重要基础[5]，但在高维多指接触和对象参数变化明显的条件下，仍然离不开大量建模和调参工作。近年来，深度强化学习提供了另一条路：通过与环境的大规模交互，策略可以在高保真仿真中逐步学得连续动作和长期稳定性，再借助域随机化等方法提升对现实差异的适应能力[3,6–9]。

尽管如此，现有成果仍主要集中在高成本平台、特定对象或附加感知条件下的手内操作研究。对于以 LEAP Hand 为代表的低成本灵巧手平台，如何围绕连续手内旋转任务构建一整套兼顾训练效率、部署可行性和后续扩展潜力的方案，仍然是一个很现实的问题[10]。和 Shadow Hand 这类高成本平台相比，LEAP Hand 结构紧凑、成本较低、容易复现实验，但也面临电机输出余量有限、关节控制精度受限、缺少高精度触觉传感以及真实接触参数难以准确标定等局限。这些硬件约束使策略在部署阶段难以依赖丰富外部感知或精确动力学模型，而更需要从关节状态、关节目标及其时间变化中恢复隐含的接触和对象信息。因此，本文采用以本体感觉历史为核心的设计思路，并在深度强化学习框架下完成连续旋转策略训练与真实部署研究。

因此，本文的价值主要体现在两个具体层面：一是验证低成本 LEAP Hand 是否能够通过仿真强化学习获得可部署的连续旋转能力；二是把任务建模、两阶段学习和 Sim2Real

部署中的关键接口整理成一条可复现实验链路，为后续扩展到更复杂对象提供基础。

1.2 国内外研究现状

在灵巧手手内操作研究中，强化学习与 Sim2Real 结合是一条重要路线。OpenAI 在 Shadow Hand 平台上完成了基于视觉的手内重定向策略学习，展示了仿真训练结合域随机化在复杂多指系统中的迁移潜力[3]。随后，OpenAI 又展示了单手解魔方任务，将长期规划、精细接触操作与零样本仿真到现实迁移结合起来，使高复杂度灵巧操作受到更多关注[8]。

从训练方法和实验平台的发展来看，PPO 由于实现简洁、训练稳定，在机器人强化学习任务中被广泛采用[11]。与此同时，Isaac Gym 这类 GPU 并行仿真平台显著提高了样本采集效率，使得在单机环境下训练大规模灵巧手策略成为可能[9]。强化学习算法与高并行仿真平台的结合，推动了灵巧操作研究从单一演示逐步走向系统化训练与评估。

具体到灵巧手学习控制，Rajeswaran 等将示范数据与深度强化学习结合，用于提升高维灵巧操作任务的学习效果[12]；Zhu 等进一步展示了在低成本硬件上通过强化学习完成灵巧操作的可能性[13]；Van Hoof 等则较早探索了触觉特征在机器人手内操作学习中的作用[14]。这些工作分别从示范利用、低成本平台和触觉感知三个角度推进了学习式灵巧操作的发展。近年来，手内旋转任务也出现了若干代表性研究。Qi 等提出 HORA，通过在圆柱体对象上训练基础策略并利用本体感觉历史进行在线适配，实现了对多种真实物体的直接旋转部署，表明基于快速自适应的两阶段方法在手内旋转任务中具有较高实用价值[15]。在此基础上，Qi 等又提出融合视觉和触觉信息的 RotateIt，将研究对象扩展到多轴旋转与多模态感知条件下的手内旋转，进一步提升了对复杂对象属性的识别与控制能力[16]。Yang 等提出 AnyRotate，利用密集触觉信息实现了重力方向变化条件下的多轴手内旋转，并展示了面向未见对象的零样本迁移能力[17]。代表性工作的横向对比如表 1.1 所示。

表 1.1 中的对比显示，HORA 已经证明两阶段快速适应路线在手内旋转任务中的价值，RotateIt 和 AnyRotate 则把研究推进到多模态感知、多轴旋转和更复杂对象泛化上。本文并不以“全面超过这些工作在任意物体泛化上的能力”为目标，而是聚焦一个更具体的缺口：在低成本 LEAP Hand、无视觉/无触觉反馈、仅依赖本体感觉历史的条件下，建立一条可复现的圆柱体连续旋转训练、评估与实机闭环链路，并分析该链路在圆柱体参数分布外条件下的主要退化因素。

国内在灵巧操作领域的研究起步相对较晚，但近年来在感知、抓取与泛化学习方面进展明显。Xia 等对灵巧操作中的感知问题进行了系统综述，指出视觉、触觉和本体感觉

表 1.1 手内旋转相关代表性工作的横向对比

工作	平台/对象	感知模态	任务特点	迁移或泛化方式
OpenAI 手内操作[3,8]	Shadow Hand，高成本平台	视觉、本体感觉等	姿态重定向、魔方操作	大规模域随机化，真实部署
HORA[15]	灵巧手，圆柱体基础对象	本体历史、训练期特权信息	单轴连续旋转	两阶段快速运动适应
RotateIt[16]	多类旋转对象	视觉、触觉、本体感觉	多轴手内旋转	多模态估计对象属性
AnyRotate[17]	多对象，重力方向变化	密集触觉、本体感觉	多轴旋转，未见对象	触觉驱动零样本迁移

的协同是提升灵巧操作能力的重要方向[4]。在学习方法方面，Yuan 等提出跨构型灵巧抓取强化学习框架，探索了统一动作表示在不同灵巧手平台之间的迁移能力，反映出国内研究已经开始从单一平台的策略学习转向更强调泛化与统一表示的方向[18]。不过，与灵巧抓取和多模态感知研究相比，直接面向连续手内旋转并强调完整 Sim2Real 部署链路的工作仍然不多。

已有研究已经把手内旋转推进到多模态感知、多轴旋转和未见对象迁移等方向，但这些成果往往依赖更昂贵的平台、更强的传感配置，或更复杂的对象状态估计。低成本灵巧手上的连续旋转仍有一个实际缺口：在视觉和触觉都不作为策略输入的情况下，系统能否仅凭本体感觉历史完成稳定闭环，并在圆柱体参数变化时保持一定鲁棒性。本文聚焦这一缺口，研究范围主要限定在圆柱体及少量近圆柱对象，不与多模态方法在任意对象泛化能力上作直接比较。

1.3 本文研究内容与主要贡献

基于上述研究现状，本文将研究重点收敛到 LEAP Hand 平台上的圆柱体连续手内旋转任务，目标是在 Isaac Lab 仿真环境中建立可训练、可评估、可部署的连续旋转系统，并在真实平台上完成验证。围绕这一目标，本文主要回答三个仍未被现有文献直接覆盖的问题。第一，HORA 类两阶段适应方法在低成本 LEAP Hand、无视觉和无触觉反馈、真实执行器余量有限的条件下，是否仍然能够形成可执行的连续旋转闭环。第二，训练期能够获得的对象位置、尺度、质量、摩擦和质心偏移等属性，应该直接作为策略主干的输入，还是更适合压缩为面向控制的低维技能先验，由本体感觉历史来估计。第三，当第二阶段部署策略在圆柱体物理参数分布外条件下出现退化时，主要风险来自哪些对象属性，教师——学生动作一致性约束和质心偏移覆盖能否缓解这类退化。

围绕这三个问题，本文按“任务建模——策略学习——部署验证”的逻辑展开。首先，结合 LEAP Hand 的运动特性和真实控制接口，建立圆柱体连续旋转任务的状态观测、动作表示、奖励函数、终止条件和评测指标，使仿真训练、统一评估与实机部署具有一致定义。其次，构建技能先验驱动的本体自适应旋转框架：第一阶段利用训练期可获得的对象物理信息学习旋转技能先验和教师策略，第二阶段在冻结动作主干的基础上，通过本体感觉历史估计同一技能先验，从而让部署策略在无法直接获得对象质量、摩擦和质心等参数的条件下仍能闭环执行。最后，针对第二阶段在分布外退化的问题，本文引入教师——部署动作一致性约束，并扩展质心偏移等风险参数的训练覆盖，令部署侧在保留低维技能先验的同时尽量贴近教师侧动作语义。

在仿真验证层面，本文完成 ID/OOD 圆柱体参数范围下的量化评测，并通过消融实验分析特权信息类型、历史长度、历史编码器结构与动作一致性损失的作用。实机部分将同一增强策略部署到真实 LEAP Hand，在标准圆柱体、尺寸变化圆柱体和若干近圆柱真实物体上进行闭环测试，用真实执行结果检查策略链路的可执行性与边界。相应地，本文的主要贡献包括：面向 LEAP Hand 连续手内旋转建立与真实部署接口一致的任务建模、奖励设计和统一评测流程；构建技能先验驱动的两阶段本体自适应框架，把训练期特权信息学习到的旋转能力转化为仅依赖本体感觉历史的可部署策略；针对第二阶段分布外退化，进一步引入动作一致性约束和风险参数覆盖，并在 ID/OOD 仿真评测及真实 LEAP Hand 闭环部署中验证其在当前任务范围内的有效性。

1.4 论文结构

全文共分为五章。第一章介绍研究背景、相关工作、本文的研究内容与论文结构。第二章围绕连续手内旋转任务进行建模，说明任务定义、状态与动作表示、奖励设计以及训练目标。第三章介绍算法设计，重点说明技能先验驱动的本体自适应旋转框架及其部署一致性设计。第四章给出实验配置、仿真评测、实机部署结果及综合分析。第五章总结全文工作，并对后续研究方向进行展望。

1.5 本章小结

本章介绍了连续手内旋转任务的研究背景、国内外研究现状、本文研究内容和论文结构。相关工作已经将手内旋转研究推进到泛化能力和真实部署阶段，但低成本灵巧手平台上的完整训练与部署链路仍不充分。基于这一判断，后续章节将围绕 LEAP Hand 连续手内旋转任务展开建模、方法设计与实验验证。

2 任务建模

2.1 连续旋转任务定义与系统状态

本文研究的是 LEAP Hand 对圆柱体的连续手内旋转控制。这里的“连续”不是把物体转到某个固定角度就结束，而是在不丢失抓持稳定性的前提下，让物体在回合持续时间内尽量保持同一旋转方向上的角运动。与一次性重定向不同，这个任务更关注一段时间内能否保持稳定、可控的角速度，因此策略目标并不是“到达某个离散姿态”，而是让旋转过程尽量平滑、持续，并避免掉落和过度控制。

为了突出连续旋转本身，本文将主要对象限定为圆柱体。圆柱体具有明确主轴和相对规则的局部接触几何，可以在降低形状复杂度的同时保留多指协同、接触切换、摩擦不确定性和抓持稳定性等关键难点。该设定也与 HORA 一类手内旋转研究的基本思路一致：先在结构规整的对象上建立可训练的旋转控制系统，再讨论向更复杂对象迁移的可能性[15]。

在仿真环境中，物体被初始化在 LEAP Hand 四指可形成稳定抓取的区域内，初始主轴与世界坐标系 z 轴基本对齐，策略随后以固定控制频率输出关节目标增量，驱动物体绕世界坐标系的 z 轴连续旋转。设离散控制时刻为 t，控制步长为 $\Delta t = 4/120\,\mathrm{s} \approx 0.0333\,\mathrm{s}$，目标旋转轴为

$$e_z = [0, 0, 1]^\top. \tag{2.1}$$

则任务可以概括为：在给定回合时长内最大化物体沿 e_z 方向的累计旋转量，同时抑制掉落、过大关节运动和不必要的控制代价。由此得到的是一个以本体感觉输入为主、带有对象物理参数随机化的连续控制问题。

图 2.1 总结了本文任务建模中最核心的变量。图中间展示了真实 LEAP Hand 抓持圆柱体的交互场景，右上角给出了长时回合中的优化目标。外层闭环则对应强化学习接口：策略输入由短时本体感觉历史堆叠而成，动作表示为 16 维关节目标增量，奖励由旋转收益、稳定性约束和控制代价共同组成，终止条件用于处理掉落或超时。需要说明的是，图中的圆柱体并不是单纯的几何简化，而是后续两阶段方法中技能先验学习的基础对象。

2.2 策略输入、特权信息与动作空间

本文采用以本体感觉为主的观测形式。单帧本体感觉观测由归一化关节位置和当前关节目标位置组成，记为

$$o_t = \left[\tilde{q}_t,\ q_t^{\mathrm{tar}}\right] \in \mathbb{R}^{32}. \tag{2.2}$$

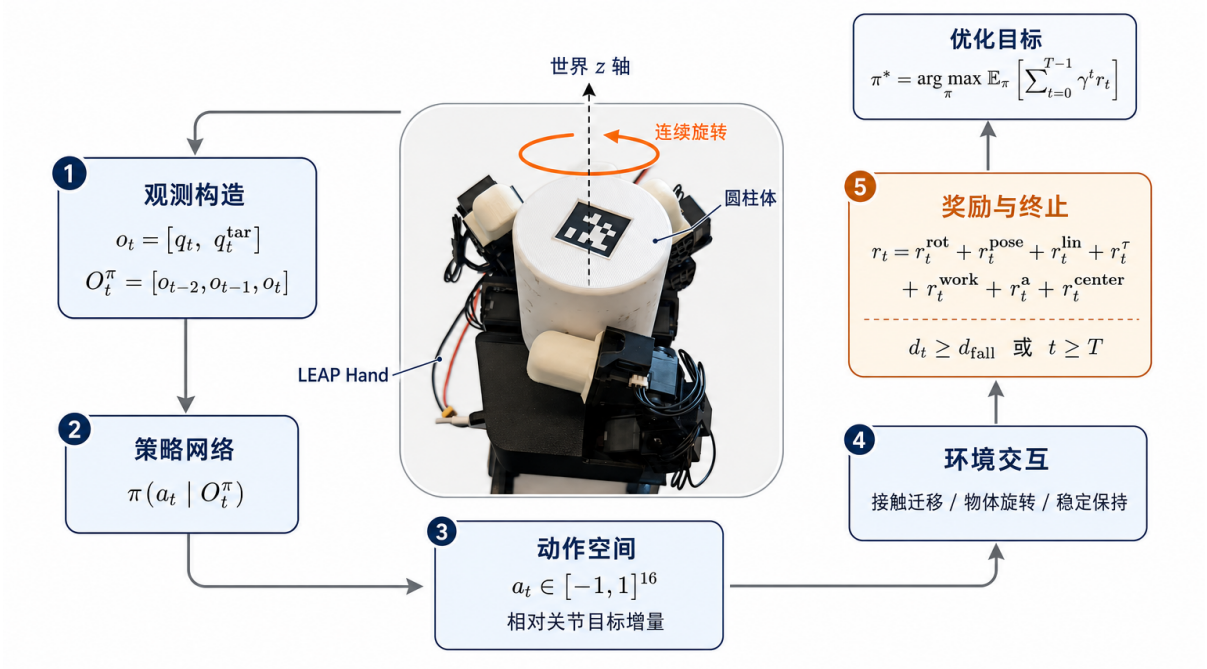

图 2.1 圆柱体连续手内旋转任务建模示意图

其中，前 16 维为归一化关节位置 $\tilde{q}_t$，后 16 维为底层位置控制器跟踪的关节目标 q_t^{tar}。关节位置归一化定义为

$$\tilde{q}_t = \frac{2q_t - q_{\max} - q_{\min}}{q_{\max} - q_{\min}}, \tag{2.3}$$

其中 $q_{\min}, q_{\max} \in \mathbb{R}^{16}$ 分别表示各关节的下界和上界。该观测不包含视觉、触觉或真实对象参数，因而与后续实机部署阶段能够直接读取的信息保持一致。

教师策略的输入采用最近 3 个控制时刻的本体感觉堆叠：

$$O_t^\pi = [o_{t-2}; o_{t-1}; o_t] \in \mathbb{R}^{96}. \tag{2.4}$$

这里的 96 维来自 3×32。短时历史可以提供接触趋势、关节变化和物体运动方向等时序信息，使策略在保持低维输入的同时仍能感知到动作后的状态变化。

为了给两阶段学习提供额外监督，本文在训练阶段引入特权信息

$$\xi_t = \left[p_t^o, \rho, m, \mu, c^\top\right]^\top \in \mathbb{R}^9. \tag{2.5}$$

其中 p_t^o 表示物体位置，ρ 为尺寸参数，m 为质量，μ 为摩擦系数，c 为质心偏移。它们在仿真中可直接读取，但在真实部署时不作为策略输入，只在第一阶段训练和第二阶段的蒸馏过程中起作用。

本文将第二阶段适应模块所用的历史窗口记为

$$h_t = [o_{t-29}; o_{t-28}; \ldots; o_t] \in \mathbb{R}^{30\times 32}. \tag{2.6}$$

这一窗口长度在后续实验中固定使用，用来让历史编码器从较长的本体感觉序列中估计对象相关隐变量。

策略输出为 16 维连续动作

$$a_t \in [-1, 1]^{16}. \tag{2.7}$$

它并不直接表示关节力矩，而是表示相对关节目标的增量。设上一控制时刻的关节目标为 q_{t-1}^{tar}，动作缩放系数为 α，则当前关节目标更新为

$$q_t^{\text{tar}} = \text{clip}\left(q_{t-1}^{\text{tar}} + \alpha a_t,\ q_{\min}, q_{\max}\right). \tag{2.8}$$

本文实现中取 $\alpha = 1/24$，这意味着在动作取边界值时，单个控制周期内的目标增量约为 $0.0417\,\text{rad}$。这种表示方式便于对接真实 LEAP Hand 的位置控制接口，也比直接输出力矩更稳定，更适合保持仿真和实机部署之间的动作语义一致。

2.3 旋转度量与奖励函数

连续手内旋转的核心评价量是物体沿目标轴的角运动。本文用相邻两个时刻的四元数差分表示物体姿态增量：

$$\Delta Q_t = Q_t^o \otimes (Q_{t-1}^o)^{-1}, \tag{2.9}$$

其中 $\otimes$ 表示四元数乘法。再将姿态增量映射为轴角向量

$$\phi_t = \text{Log}(\Delta Q_t), \tag{2.10}$$

即可得到当前控制步长内的平均角速度估计

$$\hat{\omega}_t^o = \frac{\phi_t}{\Delta t}. \tag{2.11}$$

沿目标轴的旋转速度定义为

$$\omega_t^{\parallel} = e_z^{\top} \hat{\omega}_t^o. \tag{2.12}$$

本文只奖励世界坐标系 z 轴上的净旋转，因为任务目标是围绕竖直轴持续旋转；若将倾斜或平移也计入收益，策略可能利用非目标运动获得回报，却没有形成稳定的轴向旋转。

为了避免瞬时角速度过大对训练造成干扰，旋转奖励采用截断形式：

$$r_t^{\text{rot}} = \lambda_{\text{rot}}\, \text{clip}\left(\omega_t^{\parallel},\ r_{\min}, r_{\max}\right), \tag{2.13}$$

其中 $\lambda_{\text{rot}} > 0$ 为旋转奖励权重，$r_{\min}$ 和 $r_{\max}$ 为角速度截断阈值。总奖励则写成

$$r_t = r_t^{\text{rot}} + r_t^{\text{pose}} + r_t^{\text{lin}} + r_t^{\tau} + r_t^{\text{work}} + r_t^{a} + r_t^{\text{center}}. \tag{2.14}$$

其中，姿态偏离惩罚用于抑制初始抓持姿态的大幅偏移：

$$r_t^{\text{pose}} = -\lambda_q \|q_t - q_0\|_2^2. \tag{2.15}$$

物体线速度惩罚写为

$$\hat{v}_t^o = \frac{p_t^o - p_{t-1}^o}{\Delta t}, \qquad r_t^{\text{lin}} = -\lambda_v \|\hat{v}_t^o\|_1. \tag{2.16}$$

关节力矩和关节功率惩罚分别为

$$r_t^{\tau} = -\lambda_\tau \|\tau_t\|_2^2, \tag{2.17}$$

$$r_t^{\text{work}} = -\lambda_w (\tau_t^\top \dot{q}_t)^2. \tag{2.18}$$

动作幅值惩罚定义为

$$r_t^a = -\lambda_a \|a_t\|_2^2. \tag{2.19}$$

最后，针对质心偏移带来的掉落风险，本文引入中间位置惩罚

$$d_t = \|p_t^o - p_c\|_2, \qquad r_t^{\text{center}} = -\lambda_c \left[\text{clip}\left(\frac{d_t}{d_{\text{fall}}},\, 0, 1\right)\right]^2. \tag{2.20}$$

其中 p_c 为物体中心参考位置，$d_{\text{fall}} = 0.07\,\text{m}$ 为允许的最大中心偏移距离。

这些奖励项分别对应任务中的不同约束：旋转奖励给出优化方向，姿态和线速度惩罚维持抓持稳定性，力矩、功率和动作幅值惩罚限制控制代价，中间位置惩罚主要用于降低掉落风险。通过这组分量，连续旋转目标被拆解为可学习、可调节的若干约束；各项奖励权重及角速度截断阈值的具体取值将在第 4.1 节实验设置中给出。

2.4 终止条件与优化目标

本文将物体中心偏移视为主要失败判据。若物体中心相对参考位置的距离超过阈值 d_{fall}，则认为物体已经脱离手内可控区域；若回合达到最大时长 T，则回合自然结束。终止条件写为

$$\text{done}_t = \mathbb{I}\left[d_t \geq d_{\text{fall}} \ \vee \ t \geq T\right], \tag{2.21}$$

其中 $\mathbb{I}[\cdot]$ 为指示函数。后续仿真和实机评测统一采用 20 s 的回合时长，以便不同策略之间保持一致的时间尺度。

基于上述状态、观测、动作和奖励定义，连续手内旋转任务可以写成一个离散时间马尔可夫决策过程。设策略为 $\pi(a_t|O_t^\pi)$，折扣因子为 γ，则策略优化目标为

$$\pi^* = \arg\max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r_t\right]. \tag{2.22}$$

该目标并非单纯追求更大的瞬时旋转速度，而是要求策略在整个回合内保持稳定、持续且可执行的旋转行为。本章建模结果为第三章的两阶段学习算法提供统一的任务、观测和奖励基础。

2.5 本章小结

本章对 LEAP Hand 圆柱体连续手内旋转任务进行了形式化建模，依次定义任务对象、目标旋转轴、控制步长、系统状态、策略观测、训练期特权信息和相对关节目标动作空间，并给出轴向旋转量计算方式、奖励函数、终止条件和策略优化目标。这些定义为后续算法设计和实验验证提供统一基础。

3 算法设计

3.1 技能先验驱动的本体自适应旋转框架

针对第二章建立的 LEAP Hand 圆柱体连续手内旋转任务，本文设计技能先验驱动的本体自适应旋转框架。框架并不将“教师策略”和“部署策略”处理为两套彼此独立的网络，而是把训练期对象信息、本体感觉历史和最终部署动作组织在同一动作生成链路中：第一阶段学习旋转技能先验和教师策略，第二阶段在冻结动作主干的基础上由本体感觉历史恢复同一技能先验，最终得到可直接部署的策略。

图 3.1 给出了框架的信息流。第一阶段在仿真中利用特权信息学习旋转技能；第二阶段冻结第一阶段得到的动作主干，只训练历史编码器估计同一技能先验；部署阶段移除特权信息，由本体历史和动作主干完成真实控制。这样处理的目的，是把训练期可直接读取的对象参数转化为部署期可由关节历史间接恢复的中间表示。

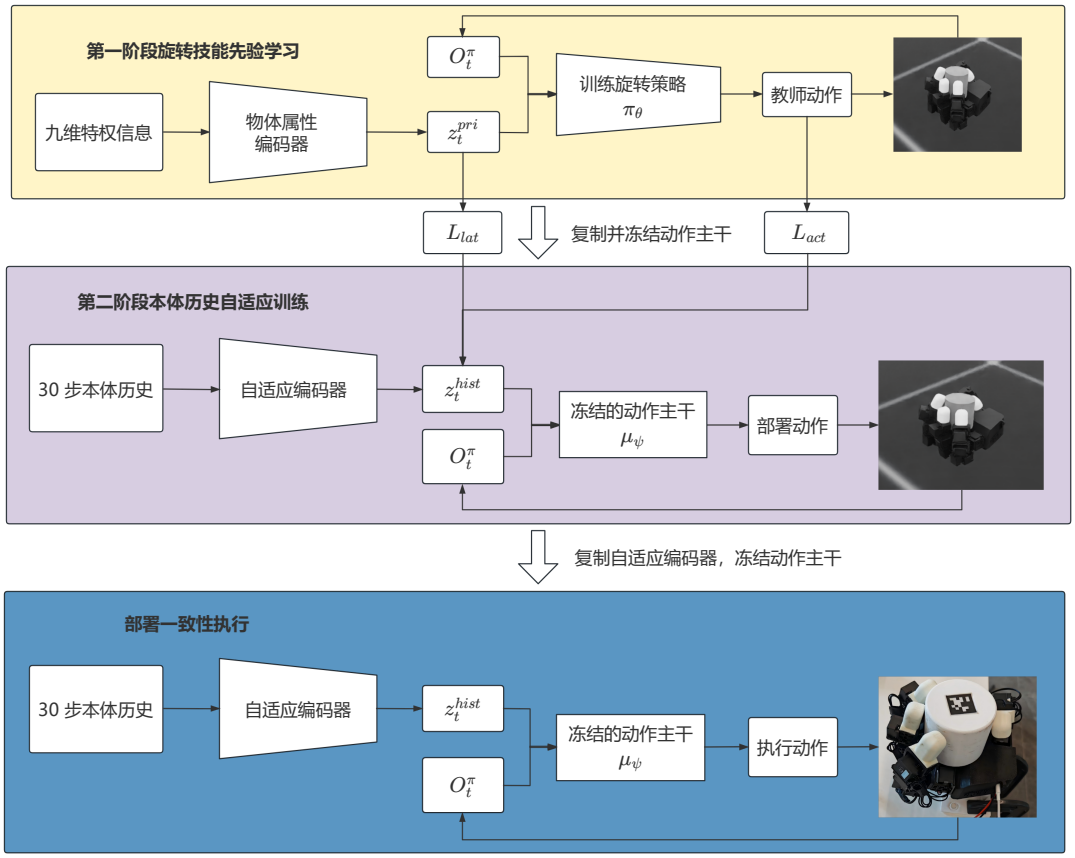

图 3.1 技能先验驱动的两阶段训练与部署执行框架

表 3.1 汇总了本文主要训练与对比的三种策略。三者共享同一动作主干，但输入和训练目标不同：Teacher 只在仿真中使用特权信息，是阶段一的参考上界；Deploy-Base

只使用本体历史，不再接触任何对象物理参数，更接近 HORA/RMA 的原始部署思路；Deploy-Refined 则在 Deploy-Base 的基础上加入动作一致性约束，并配合更宽的质心偏移分布，以缓解分布外退化。

表 3.1 三种策略的输入、训练目标与部署属性

策略	本体历史	特权信息	训练目标	训练分布	可实机部署
Teacher	短历史观测	使用	PPO 强化学习	基础圆柱参数分布	否
Deploy-Base	使用	不使用	L_{lat}	基础圆柱参数分布	是
Deploy-Refined	使用	不使用	$L_{\text{lat}} + \lambda_{\text{act}} L_{\text{act}}$	更宽质心偏移分布	是

从动作接口看，上述三种策略的差别主要在技能先验的来源，而不是控制量的定义。统一写为

$$z_t = \begin{cases} z_t^{\text{pri}} = E_{\text{pri}}(\xi_t), & \text{教师侧先验,} \\ z_t^{\text{hist}} = E_{\text{hist}}(h_t), & \text{部署侧先验,} \end{cases} \qquad \mu_t = \mu_\psi([O_t^\pi; z_t]). \tag{3.1}$$

第一阶段训练时，策略围绕 μ_t 构造连续动作分布并采样动作；第二阶段监督和真实部署时，则使用截断后的确定性均值动作。

3.2 第一阶段：旋转技能先验学习

第一阶段的任务是在仿真中先学习稳定的旋转控制能力，获得一个具有较强旋转能力的教师侧控制策略。为此，本文将训练期能够直接读取的对象属性定义为特权信息

$$\xi_t = [p_t^o, \rho, m, \mu, c^\top]^\top \in \mathbb{R}^9, \tag{3.2}$$

其中 p_t^o 是物体位置，ρ 是尺寸参数，m 是质量，μ 是摩擦系数，c 是质心偏移。随后，特权信息被编码为低维技能先验

$$z_t^{\text{pri}} = E_{\text{pri}}(\xi_t), \qquad z_t^{\text{pri}} \in \mathbb{R}^8. \tag{3.3}$$

这里的 8 维并不是简单压缩，而是刻意把对象属性映射到更贴近控制需求的中间空间：一方面减少策略直接记忆原始物理量的倾向，另一方面也给第二阶段留下一个清晰的监督目标。表 4.6 给出的直接特权输入消融结果进一步说明，原始物理参数直接拼接并没有自动带来更好的教师能力或部署效果，低维技能先验在当前任务中起到了表征约束作用。

将本体观测与技能先验拼接后，第一阶段策略输入写为

$$\bar{O}_t = [O_t^\pi; z_t^{\text{pri}}]. \tag{3.4}$$

在 PPO 训练中，策略网络输出动作分布，价值网络估计回报，策略比率定义为

$$r_t(\theta) = \frac{\pi_\theta(a_t|\bar{O}_t)}{\pi_{\theta_{\text{old}}}(a_t|\bar{O}_t)}. \tag{3.5}$$

PPO 的裁剪目标为

$$L_{\text{clip}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big], \tag{3.6}$$

其中 ϵ 为裁剪系数。综合价值函数损失和熵正则后，第一阶段目标写为

$$\max_\theta\ J_{\text{stage1}}(\theta) = L_{\text{clip}}(\theta) - \lambda_V L_V(\theta) + \lambda_H H(\pi_\theta), \tag{3.7}$$

其中 $L_V(\theta)$ 为价值函数损失，$H(\pi_\theta)$ 为策略熵项。

算法 3.1 给出了这一阶段的训练流程。实现上，本文把第一阶段分成轨迹采样和小批量 PPO 更新两个层面：外层并行环境负责收集教师侧数据，内层则在固定采样批次上迭代更新特权信息编码器、策略网络和价值网络。这样做的好处是，训练数据和优化目标都围绕同一套技能先验展开，后面第二阶段学习本体历史时就有了明确的参照。

算法 3.1：第一阶段旋转技能先验学习

输入： 并行环境 $\mathcal{E}$，策略观测 O_t^π，特权信息 ξ_t，奖励 r_t

超参数： 采样步长 H，外层迭代次数 K_1，PPO 更新轮数 M_1，小批量大小 B_1

输出： 特权信息编码器 E_{pri}，策略 π_θ，价值函数 V_ϕ

初始化： E_{pri}、π_θ、V_ϕ，并令 $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$

for $k = 1, \ldots, K_1$ **do**
  $\mathcal{D} \leftarrow \emptyset$
  for $t = 0, \ldots, H-1$ **do**
    $z_t^{\text{pri}} \leftarrow E_{\text{pri}}(\xi_t)$
    $\bar{O}_t \leftarrow [O_t^\pi; z_t^{\text{pri}}]$
    $a_t \sim \pi_{\theta_{\text{old}}}(\cdot|\bar{O}_t)$，$p_t^{\text{old}} \leftarrow \pi_{\theta_{\text{old}}}(a_t|\bar{O}_t)$
    $v_t \leftarrow V_\phi(\bar{O}_t)$
    $(r_t, O_{t+1}^\pi, \xi_{t+1}) \leftarrow \mathcal{E}(a_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\bar{O}_t, a_t, r_t, v_t, p_t^{\text{old}})\}$
  根据 $\mathcal{D}$ 计算回报 $\hat{G}_t$ 与优势 $\hat{A}_t$
  for $m = 1, \ldots, M_1$ **do**
    采样小批量 $\mathcal{B} \subset \mathcal{D}$，并用保存的 p_t^{old} 计算策略比率
    按式 (3.5) 和式 (3.7) 计算 PPO 更新目标
    对 E_{pri}、π_θ、V_ϕ 执行 Adam 更新
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
返回 $E_{\text{pri}}, \pi_\theta, V_\phi$

因此，第一阶段得到的是带特权信息的教师侧旋转策略，而不是最终部署策略。它的主要作用是形成稳定动作主干和可监督的技能先验空间，使第二阶段能够在同一动作接口下学习本体历史适应模块。

3.3 第二阶段：本体历史自适应训练

第一阶段用到了特权信息，但这部分信息在真实部署时并不存在。因此，第二阶段必须把“依赖对象参数”的能力，转成“只看本体历史也能估计”的能力。为此，本文冻结第一阶段已经学到的动作主干，仅训练历史编码器 $E_{\text{hist}}(\cdot)$，让系统从本体感觉历史中恢复和教师侧一致的技能先验。设第二章定义的本体历史为

$$h_t = [o_{t-29}; o_{t-28}; \ldots; o_t] \in \mathbb{R}^{30\times 32}, \tag{3.8}$$

则部署侧技能先验写为

$$z_t^{\text{hist}} = E_{\text{hist}}(h_t), \qquad z_t^{\text{hist}} \in \mathbb{R}^8. \tag{3.9}$$

本文采用固定长度的 30 步本体感觉历史作为适应输入。在约 30 Hz 的控制频率下，该窗口覆盖约 1 s 的近期关节响应，既包含若干次动作更新后的动态线索，又不会使真实部署端缓存过长。表 4.7 显示，历史长度继续增加并不带来单调收益，因此 30 帧主要是历史覆盖、旋转效率和部署开销之间的折中。

历史编码器采用一维时序卷积结构。具体而言，单帧 32 维本体观测先经过两层线性映射和 ReLU 激活得到通道特征，随后沿时间维依次通过三层一维卷积进行聚合，最后展平并投影到 8 维技能先验空间。相较于直接展平历史窗口，卷积结构能够显式利用相邻时刻的局部变化；与递归结构相比，它在部署端不需要维护额外隐状态。表 4.8 给出的结构消融也支持这一选择。

第二阶段只训练历史编码器，同时冻结动作主干，原因主要有三点。其一，冻结动作主干可以保留第一阶段已经形成的稳定动作生成方式，降低监督训练破坏原有旋转行为的风险。其二，将适应问题限制在 8 维技能先验空间内，可以降低从本体历史到动作的学习难度。其三，教师侧和部署侧共享同一动作接口，便于后续直接比较两者的输出差异。

表 3.1 里的 Deploy-Base 可视为更接近 HORA/RMA 原始部署思路的对照基线：它仅通过特征对齐完成从教师侧到部署侧的迁移，不额外引入动作一致性约束或更宽的质心偏移分布。与之相比，Deploy-Refined 在保留这一基本框架的基础上进一步加入动作一致性约束，并配合更宽的质心偏移分布，以减轻 OOD 条件下的性能退化。

第二阶段的教师侧与部署侧动作均由同一冻结后的动作主干产生，分别记为

$$\mu_t^{\text{tea}} = \mu_\psi\Big([O_t^\pi; z_t^{\text{pri}}]\Big), \qquad \mu_t^{\text{dep}} = \mu_\psi\big([O_t^\pi; z_t^{\text{hist}}]\big). \tag{3.10}$$

第一项监督历史编码器恢复技能先验，因此定义潜变量一致性损失

$$L_{\text{lat}} = \left\| z_t^{\text{hist}} - z_t^{\text{pri}} \right\|_2^2, \tag{3.11}$$

第二项则直接对齐教师和部署动作

$$L_{\text{act}} = \left\| \text{clip}(\mu_t^{\text{dep}}, -1, 1) - \text{clip}(\mu_t^{\text{tea}}, -1, 1) \right\|_2^2. \tag{3.12}$$

于是第二阶段总损失为

$$L_{\text{stage2}} = L_{\text{lat}} + \lambda_{\text{act}} L_{\text{act}}, \tag{3.13}$$

其中 λ_{act} 为动作一致性权重。前者保证历史编码器学到的表示和教师侧技能先验保持接近，后者则进一步约束最终动作输出，避免部署侧在未见参数条件下出现明显偏移。

与第一阶段不同，第二阶段的训练数据并不是离线固定数据集，而是由部署侧闭环采样得到。每一步都由策略与环境交互，再用新观测更新历史窗口 h_t，因此历史编码器实际上是在一个与真实部署更接近的闭环条件下被训练出来的。为了处理 OOD 退化，本文又将对象参数分布扩展为更宽的风险分布 $p_{\text{risk}}(\eta)$，使历史编码器在训练时能见到更多质心偏移样本。相应地，基础策略和增强策略在训练分布上的差别写成

$$\eta \sim p_{\text{risk}}(\eta), \qquad p_{\text{risk}}(c) \supset p_{\text{base}}(c). \tag{3.14}$$

这一步并不是重新定义任务，而是把第二阶段的适应压力显式推到更容易失真的对象参数上。

算法 3.2 给出了第二阶段训练流程。冻结动作主干后，训练过程只更新历史编码器，并在每个闭环采样周期中同时计算潜变量一致性和动作一致性两项损失。这样既保留了第一阶段形成的动作模式，也使部署侧输出能够在监督信号下向教师侧动作靠近。

算法 3.2：第二阶段面向风险对齐的本体历史自适应训练

输入： 向量化环境 $\mathcal{E}$，策略观测 O_t^{π}，本体感觉历史 h_t，特权信息 ξ_t

超参数： 对象参数分布 $p(\eta)$，更新步数 K_2，动作一致性权重 λ_{act}，学习率 β

输出： 历史编码器 E_{hist}

冻结： E_{pri} 与 μ_{ψ}

初始化： E_{hist}，Adam 优化器

采样 $\eta \sim p(\eta)$ 并重置环境，初始化 O_t^{π}、h_t 与 ξ_t

for $k = 1, \ldots, K_2$ **do**

$z_t^{\text{pri}} \leftarrow E_{\text{pri}}(\xi_t)$

$z_t^{\text{hist}} \leftarrow E_{\text{hist}}(h_t)$

$\mu_t^{\text{tea}} \leftarrow \mu_{\psi}([O_t^{\pi}; z_t^{\text{pri}}])$

$\mu_t^{\text{dep}} \leftarrow \mu_{\psi}([O_t^{\pi}; z_t^{\text{hist}}])$

$\tilde{\mu}_t^{\text{tea}} \leftarrow \text{clip}(\mu_t^{\text{tea}}, -1, 1)$

$\tilde{\mu}_t^{\text{dep}} \leftarrow \text{clip}(\mu_t^{\text{dep}}, -1, 1)$

$L_{\text{lat}} \leftarrow \text{mean}\, \|z_t^{\text{hist}} - z_t^{\text{pri}}\|_2^2$

$L_{\text{act}} \leftarrow \text{mean}\, \|\tilde{\mu}_t^{\text{dep}} - \tilde{\mu}_t^{\text{tea}}\|_2^2$

$L_{\text{stage2}} \leftarrow L_{\text{lat}} + \lambda_{\text{act}} L_{\text{act}}$

对 E_{hist} 执行 Adam 更新，使 L_{stage2} 最小化

执行部署侧动作 $a_t \leftarrow \tilde{\mu}_t^{\text{dep}}$，得到新观测 O_{t+1}^{π}、h_{t+1} 与终止标志

若回合结束，则重新采样 $\eta \sim p(\eta)$ 并重置环境

返回 E_{hist}

3.4 部署执行阶段

部署阶段不再做策略训练，而是把第二阶段得到的历史编码器直接用于仿真评测和真实 LEAP Hand 部署。系统在这一阶段只保留真实平台可直接读取的本体感觉观测 O_t^{π} 和历史窗口 h_t，再由历史编码器估计技能先验并输出动作：

$$a_t^{\text{dep}} = \text{clip}(\mu_{\psi}([O_t^{\pi}; E_{\text{hist}}(h_t)]), -1, 1). \tag{3.15}$$

随后动作通过第二章定义的关节目标更新方式送入底层位置控制器。

真实部署时，系统先执行短时 warmup，缓慢靠近初始抓取姿态，再用当前关节位置和关节目标构造历史窗口。随后控制环以 30 Hz 运行，历史编码器连续更新，策略每一步输出相对关节目标增量，由底层控制器完成具体驱动。为降低启动冲击，本文在初始抓取阶段保留较小闭合系数，并采用保守的电机增益设置；这一设置优先保证真实平台闭环稳定，再讨论旋转效率。

从控制意义上看，部署阶段本质上是把第二阶段学到的历史适应能力转成真实系统

上的连续控制能力。只要历史编码器能稳定恢复技能先验，部署策略就可以在不依赖对象质量、摩擦和质心参数的条件下完成动作推理。这也是本文在第四章中能够直接用同一策略做仿真评测和实机验证的原因。

3.5 本章小结

本章提出了技能先验驱动的两阶段本体自适应旋转框架。第一阶段利用仿真中的特权信息学习旋转技能先验和教师策略，第二阶段冻结动作主干，只训练历史编码器从本体感觉历史中恢复同一技能先验，并通过动作一致性约束和更宽的风险参数分布提升部署侧鲁棒性。最后，本文给出了部署执行阶段的控制方式，使仿真训练、统一评测和真实部署共享同一动作接口。该框架为第四章的仿真评测和实机验证提供了直接的方法基础。

4 实验与分析

4.1 实验设置

本章围绕四个层面展开：先验证第三章提出的两阶段旋转框架能否在圆柱体连续手内旋转任务上形成稳定技能，再看仅依赖本体感觉历史的可部署策略能否在参数变化下保持性能，接着检验面向 OOD 风险设计的训练增强是否真的缓解退化，最后把同一策略放到真实 LEAP Hand 上，观察它在执行器响应、串口通信、接触摩擦和对象几何差异同时存在时能走到什么边界。

本章的评测分为仿真训练与评测、实机部署与评估两部分。仿真侧在 Isaac Lab 中搭建 LEAP Hand 与圆柱体交互环境，任务命名为 Isaac-CylinderRotation-Leap。该平台沿用 Orbit 等机器人学习框架在模块化任务建模和并行交互上的思路，便于统一组织训练、评测和部署配置[20]。实机侧则直接使用 16 自由度 LEAP Hand，通过 Dynamixel 串口读取关节状态并发送目标位置命令，运行第二阶段得到的可部署策略。与仿真相比，真实部署还要逐项对齐观测构造、历史缓存、归一化统计量、动作尺度、关节限位和电机控制参数，同时面对执行器滞后、通信抖动以及接触物理差异带来的 Sim2Real 偏差。

第二章已经给出了任务定义、观测空间、动作空间和奖励函数的数学表达式，这里则再进一步给出仿真训练时使用到的其它重要参数：仿真环境物理步频为 120 Hz，策略每 4 个物理步输出一次动作，对应约 30 Hz 的控制频率；动作尺度为 $1/24$，训练与评测统一采用 20 s 最大回合时长，正式评测每组统计 128 个回合。策略主干使用 3 帧短历史，历史编码器使用 30 帧本体感觉历史估计 8 维技能先验。奖励函数沿用第二章定义，具体参数设置为旋转奖励权重 $\lambda_{\text{rot}} = 1.0$，角速度投影截断范围为 $[-0.5, 0.5]$；姿态偏差、物体线速度、关节力矩、关节瞬时功率、动作幅值和中心偏移惩罚权重分别为 $\lambda_q = 0.1$、$\lambda_v = 0.3$、$\lambda_\tau = 0.0005$、$\lambda_w = 0.001$、$\lambda_a = 0.00005$ 和 $\lambda_c = 0.4$，掉落判定阈值为 $d_{\text{fall}} = 0.07\,\text{m}$。

训练配置方面，第一阶段教师策略的网络结构与 PPO 主要超参数如表 4.1 所示。特权信息编码器将 9 维特权信息映射为 8 维技能先验；动作主干和价值估计均为三层 MLP，隐藏层宽度依次为 512、256、128，激活函数为 ELU。策略标准差采用固定参数形式，动作均值在评测和部署侧均裁剪到 $[-1, 1]$。

第二阶段训练冻结第一阶段动作主干，仅更新历史编码器。基础策略采用默认圆柱体参数分布和 $\lambda_{\text{act}} = 0$，增强策略在第二阶段中放宽质心偏移范围并采用 $\lambda_{\text{act}} = 1.0$ 的动作一致性损失。第二阶段正式训练使用 4096 个并行环境，训练总环境步数约为 8×10^6，历史编码器 Adam 学习率为 3×10^{-4}。

为定量评估策略对圆柱体物理属性变化的泛化性，本文将评测条件划分为分布内（In-

表 4.1 第一阶段网络结构与 PPO 训练超参数

项目	设置
并行环境数	8192
策略观测/特权信息维度	96/9
特权信息编码器	MLP$(9, 256, 128, 8)$，输出经 tanh 得到技能先验
动作主干与价值网络	MLP$(104, 512, 256, 128)$，ELU 激活，连续动作高斯策略
策略方差	固定 log-std 参数，初始化为 -1.0
采样步长与小批量大小	96 步，mini-batch 为 3072
PPO 更新轮数与最大迭代	每轮 5 个 mini-epoch，最大 5000 个 epoch
学习率与调度	5×10^{-5}，自适应学习率调度
折扣因子与 GAE 参数	$\gamma = 0.99,\ \tau = 0.95$
PPO 剪切系数与熵系数	$\epsilon = 0.2,\ \lambda_H = 0.0001$
价值损失权重与梯度裁剪	critic 系数 4，梯度范数上限 1.0

Distribution, ID）与分布外（Out-of-Distribution, OOD）两类。ID 使用与训练一致的圆柱体参数范围；OOD 在保持对象类别仍为圆柱体的前提下，扩大尺度、质量、摩擦和质心偏移范围，用于检验策略在未见物理参数条件下的表现。具体参数范围如表 4.2 所示。

表 4.2 ID 与 OOD 圆柱体参数范围

参数	ID 范围	OOD 范围
尺度系数	$[0.85, 1.15]$	$[0.80, 1.20]$
质量	$[0.03, 0.20]$ kg	$[0.02, 0.24]$ kg
摩擦系数	$[0.3, 3.0]$	$[0.2, 3.5]$
质心偏移	$[-0.01, 0.01]$ m	$[-0.015, 0.015]$ m

第三章已经给出了 Teacher、Deploy-Base 和 Deploy-Refined 的结构差异，本章评测沿用这一划分，重点比较两类差距：其一，Teacher 与可部署策略之间的差距，用来衡量从特权信息先验切换到本体历史先验后带来的部署损失；其二，Deploy-Base 与 Deploy-Refined 之间的差距，用来判断动作一致性约束和更宽质心训练分布是否缓解常规圆柱体 OOD 条件下的退化。需要注意的是，Teacher 只作为仿真参考上界，不对应真实部署；Deploy-Refined 的质心训练范围与常规 OOD 评测范围相衔接，因此这里考察的是圆柱体参数外推中的补偿效果，而不是极端质心偏移或任意对象泛化问题的彻底解决。

仿真评测的成功定义为回合未因掉落提前终止，且净旋转角达到至少一圈。本文评测结果均基于固定 checkpoint，在随机初始化的 128 个回合上统计得到，随机种子固定为 42。除成功率外，本文还统计以下指标：存活时间为回合实际持续时间除以 20 s 上限后得

到的归一化值；净旋转圈数由相邻两帧物体四元数差分得到的轴向旋转角累加后除以 2π 得到；轴向平均角速度为存活期间未裁剪轴向角速度的时间平均值，单位为 rad/s；物体平均线速度为位置差分估计速度的 L_1 范数均值，正文表格中换算为 cm/s；平均关节命令力矩为仿真中控制器计算力矩的 L_1 范数均值。线速度和命令力矩越低通常表示控制过程越平稳，其余指标越高表示任务完成效果越好。需要说明的是，本文保留训练过程中的平均回报曲线用于说明单次训练的优化趋势；真正用于比较策略稳定性和泛化能力的是固定 checkpoint 在 128 个随机初始回合上的评测表和对比图。

4.2 仿真训练与评测

本小节从仿真训练收敛、单回合行为和参数泛化三个角度分析两阶段算法的效果。首先需要确认第一阶段教师策略是否已经在当前任务建模、奖励设计和训练配置下形成稳定、持续的旋转技能。Teacher 执行时依赖训练阶段特权信息，并不对应真实系统中的最终部署形式，但它提供了两阶段方法的技能上界，也是第二阶段适应训练的教师来源。只有第一阶段具备可靠旋转能力，后续关于本体历史适应和参数泛化的比较才有明确基准。

图 4.1 给出了第一阶段教师策略的向量化 PPO 平均回报曲线。该曲线记录正式训练中并行环境批次的平均回报，用于观察优化过程的走向。曲线前段由负转正，随后持续抬升，并在后期稳定在较高水平，说明第三章的奖励设计能够为连续旋转提供清晰的优化信号，教师策略也从早期随机探索逐渐过渡到稳定的持续旋转。后文关于有效性和泛化性的判断，主要依据表 4.3 中 128 个随机初始回合的固定 checkpoint 评测。

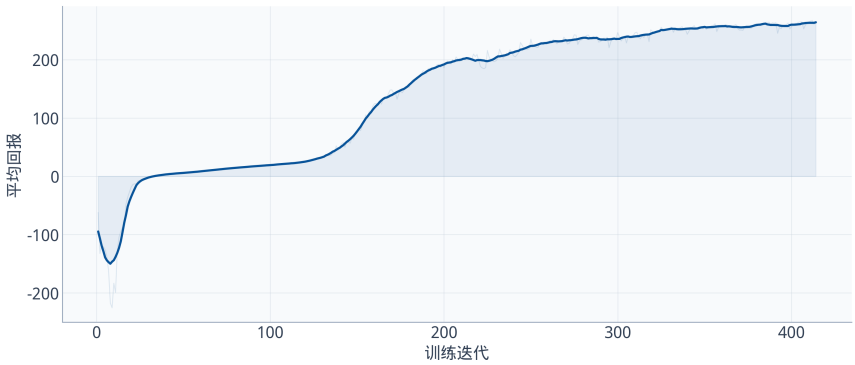

图 4.1 第一阶段教师策略向量化 PPO 训练平均回报曲线

平均回报只能反映总体优化趋势，无法单独说明策略学到的行为类型。图 4.2 因此补充给出四个核心指标。轴向平均角速度在训练后期稳定到约 1.2 rad/s，对应持续绕目标轴产生的有效角运动，而非偶发姿态抖动。物体平均线速度始终较低，说明旋转收益并非来自剧烈甩动或大幅平移。掉落率在中后期快速降至接近零，实际回合时长也逐步逼近

20 s 上限，说明抓持稳定性和持续执行能力同步提高。

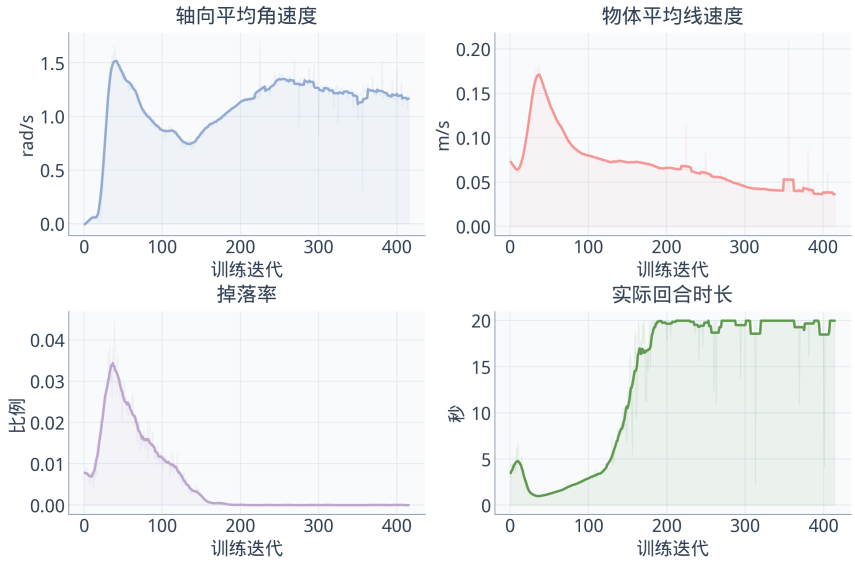

图 4.2 第一阶段四个核心训练指标曲线

图 4.3 进一步展示了第一阶段教师策略在单回合执行中的关键帧。图中选取四个时间节点，并从不同视角给出对应时刻的手内旋转状态，同时标出了帧号和时间，便于对照。可以看到，策略在连续执行过程中始终维持住对圆柱体的包络，没有出现中途脱手或明显失稳，这与前面的高奖励和低掉落率是一致的。

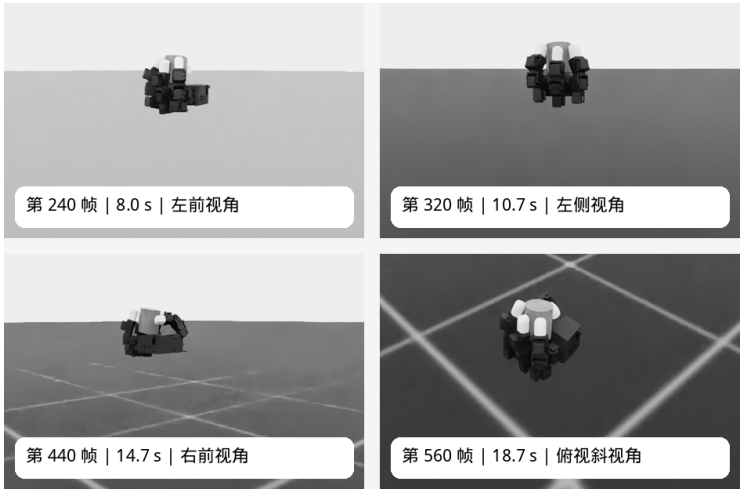

图 4.3 第一阶段教师策略在不同时间节点与不同观察视角下的关键帧

结合图 4.1、图 4.2 和图 4.3，第一阶段教师策略已经形成稳定、连续且可复用的旋转技能先验。第一阶段验证的不只是训练收敛，还包括单回合播放层面的行为可解释性，由此为第二阶段可部署策略的训练与分析提供基础。

在第一阶段获得稳定教师先验后，第二阶段只使用本体感觉历史训练可部署策略。原始第二阶段策略在 OOD 条件下仍出现明显退化，说明其对未见参数组合的外推并不充

分。针对这一问题，本文加入定向增强，并将增强前策略记为 Deploy-Base，增强后策略记为 Deploy-Refined。

对于第二阶段基策略 Deploy-Base，由于训练时冻结教师动作主干，只更新历史编码器，总奖励波动比第一阶段更明显。图 4.4 展示了该策略的总奖励曲线：尽管存在局部起伏，平滑趋势仍从低回报逐步抬升，并最终进入与第一阶段后期接近的高回报区间。也就是说，仅保留可部署观测后，第二阶段策略已经能够形成可执行的连续旋转。

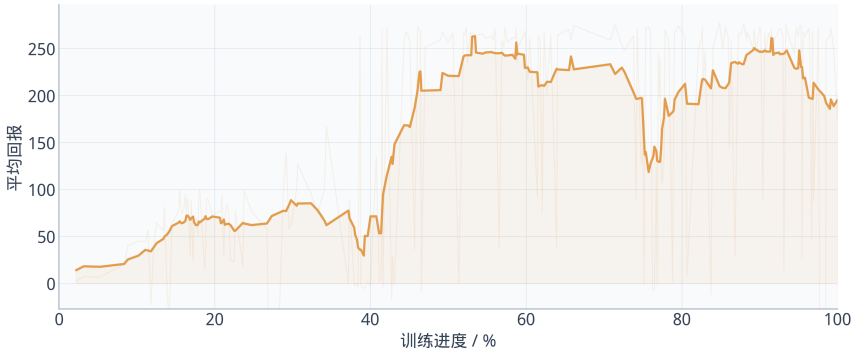

图 4.4 第二阶段基策略总奖励曲线

图 4.5 进一步给出了 Deploy-Base 的四个核心训练指标。从图中可以看到，轴向平均角速度持续上升并在训练后期稳定在约 1.1 rad/s 附近，物体平均线速度逐步回落到较低水平，表明策略已经能够在较为受控的手内区域建立持续旋转。与此同时，实际回合时长不断延长并接近 20 s 上限，潜变量重建误差也显著下降，表明历史编码器对教师技能先验的恢复在训练过程中逐步趋于稳定。

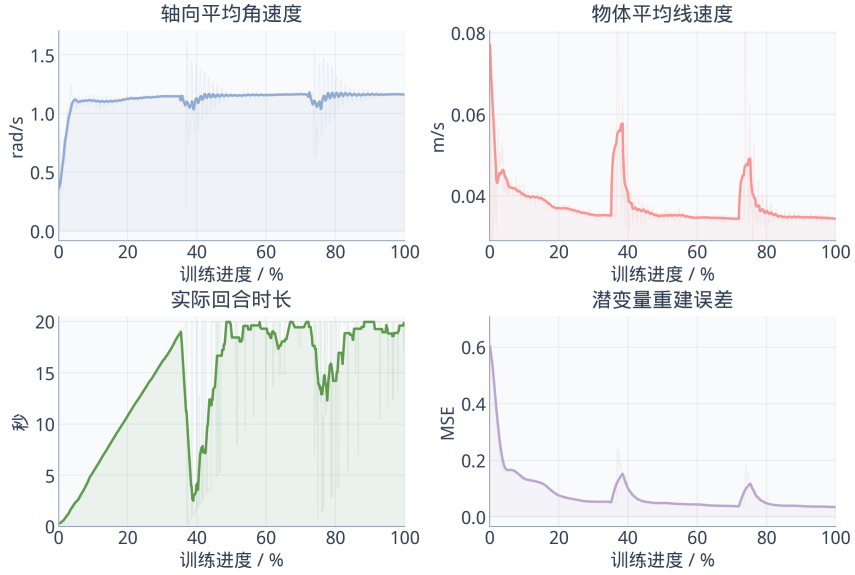

图 4.5 第二阶段基策略四个核心训练指标曲线

图 4.6 进一步给出了第二阶段基策略在单回合执行过程中的成功关键帧。图中同样选取了四个时间节点，并分别从不同观察视角展示对应时刻的手内旋转状态。从该组关键帧可以看到，Deploy-Base 在标准圆柱体样本上能够较长时间保持对物体的稳定包络，并完成连续旋转过程。该结果表明，前述训练曲线与核心指标反映的收敛结果并非仅停留在统计量层面，而是能够在具体回合播放中体现为持续的旋转执行。

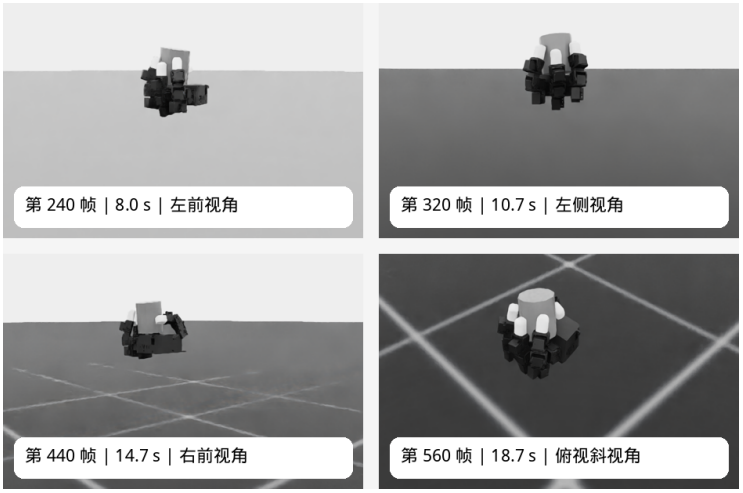

图 4.6　第二阶段基策略在不同时间节点与不同观察视角下的关键帧

图 4.4、图 4.5 与图 4.6 共同说明，第二阶段可部署基策略已经在训练收敛和单回合播放两个层面得到验证，能够依赖本体感觉历史完成较稳定的连续旋转执行。该结果也为后续分析其在偏置圆柱体条件下的退化现象提供了对比基线。

当评测对象换成更难的偏置圆柱体时，Deploy-Base 的退化更明显。图 4.7 的关键帧显示，策略在回合前段可以建立旋转，但随着执行推进，物体逐渐离开稳定接触区域，后段旋转连续性减弱，最终被挤出手内。该现象说明，仅依靠第二阶段原始训练分布和潜变量拟合目标，仍不足以覆盖偏置样本带来的接触变化。表 4.5 的特权信息类别消融也从另一侧说明，质心偏移对当前任务的接触稳定性影响更强，因此后续需要围绕偏置风险进行定向增强。

基于此，后续评测将具体检验第三章提出的第二阶段增强能否在相同部署接口下改善 OOD 表现。具体来说，增强策略对应两处训练侧调整：放宽质心偏移样本覆盖，并加入教师——部署动作一致性约束。下面先比较三类策略在 ID/OOD 条件下的整体指标，再通过消融实验区分两项增强各自的作用。

图 4.8 对比了 Teacher、Deploy-Base 和 Deploy-Refined 在 ID 与 OOD 条件下的四项核心指标。ID 条件下，两条可部署策略已经接近 Teacher，说明历史编码器基本恢复了教师侧技能先验。OOD 条件下，Deploy-Base 与教师之间的差距明显拉大，而 Deploy-Refined 在成功率、存活时间、净旋转圈数和轴向平均角速度上均有所恢复。该结果说明，第二阶

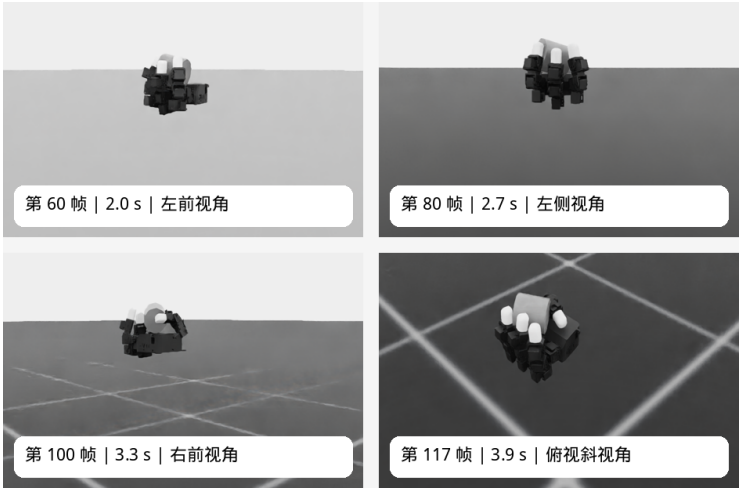

图 4.7 第二阶段基策略在较大质心偏移条件下的关键帧

段增强在不削弱分布内性能的前提下，缓解了参数外推带来的下降。

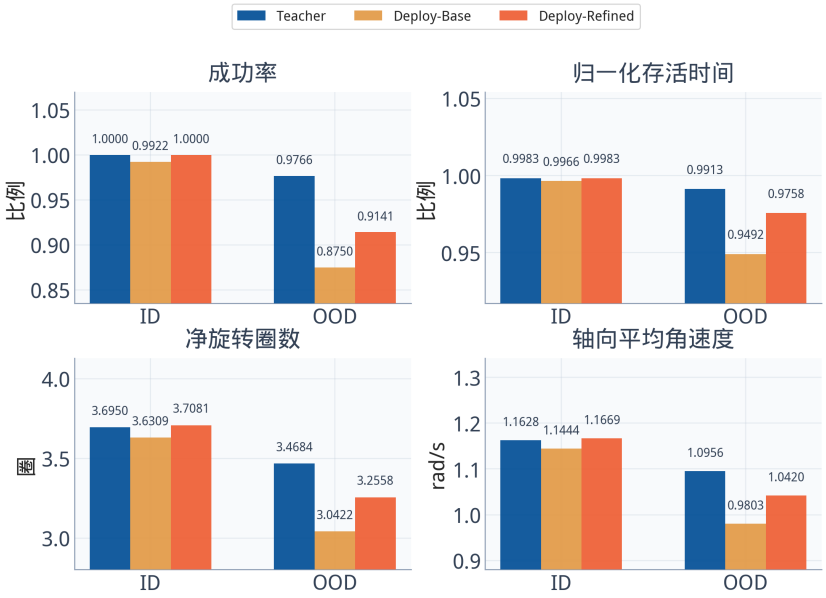

图 4.8 三种策略在 ID 和 OOD 条件下的核心评测指标对比

图 4.9 进一步给出了增强策略在较大质心偏移条件下的关键帧。与第二阶段基策略相比，Deploy-Refined 在部分时段能够维持更长的受控接触并延缓明显失稳现象，表明动作一致性监督主导下的第二阶段增强在高风险样本上具有改善效果。结合统一评测主图与关键帧诊断，该增强的主要价值在于改善常规参数外推条件下的旋转表现，并为后续继续提升极端风险条件下的稳定性提供基础。

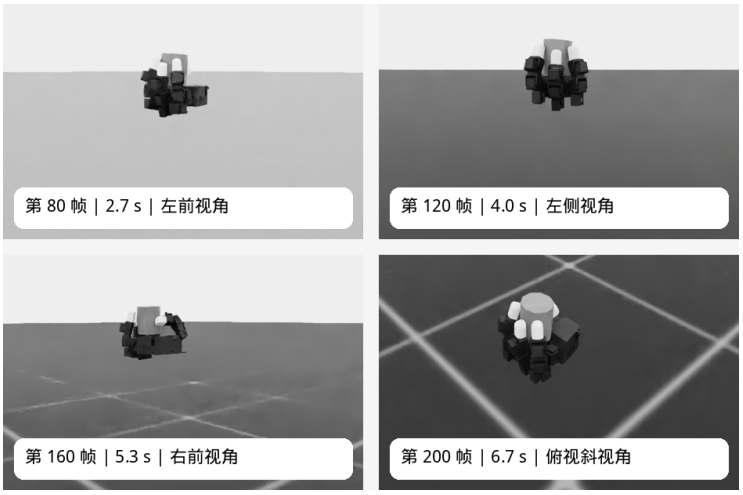

图 4.9 改进第二阶段策略在较大质心偏移条件下的关键帧

表 4.3 汇总了本文仿真评测部分最关键的定量结果。与图 4.8 相比，该表额外列出了物体平均线速度和平均关节命令力矩，因而能够同时比较任务完成率、任务持续性、旋转效率和控制平稳性。

表 4.3 三种策略在 ID 和 OOD 条件下的核心评测结果

策略	ID						OOD					
	成功率↑	存活↑	圈数↑	角速度↑	线速度↓	力矩↓	成功率↑	存活↑	圈数↑	角速度↑	线速度↓	力矩↓
教师策略	**1.0000**	**0.9983**	3.6950	1.1628	**3.4478**	2.3407	**0.9766**	**0.9913**	**3.4684**	**1.0956**	**3.9620**	2.4103
部署基线	0.9922	0.9966	3.6309	1.1444	3.5378	2.3575	0.8750	0.9492	3.0422	0.9803	4.4935	2.4433
增强部署	**1.0000**	**0.9983**	**3.7081**	**1.1669**	3.6543	**2.3373**	0.9141	0.9758	3.2558	1.0420	4.2818	**2.4087**

注：存活为归一化存活时间；角速度单位为 rad/s，线速度单位为 cm/s，力矩为仿真关节命令力矩 L_1 均值。

在 ID 条件下，三种策略都达到或接近满成功率，圆柱体训练分布内的连续旋转已经较为稳定。Deploy-Refined 的 ID 成功率为 1.0000，轴向平均角速度为 1.1669 rad/s，平均力矩为 2.3373，在保持可部署观测约束的同时取得了与教师策略相当的旋转效率和控制代价。表中 Deploy-Refined 在角速度和力矩等少数 ID 指标上略优于 Teacher，主要与有限评测回合下的对象参数、初始状态采样差异以及动作一致性训练带来的局部动作平滑有关，不能简单理解为部署策略整体超过了带特权信息的教师策略。由此可判断，动作一致性约束和更宽质心分布没有削弱分布内性能。

从 OOD 结果看，Deploy-Base 的成功率由 ID 条件下的 0.9922 降至 0.8750，轴向平均角速度降至 0.9803 rad/s，并伴随更高的物体线速度和力矩，表明仅依赖潜变量重建的第二阶段基策略在未见参数组合下稳定性不足。Deploy-Refined 将 OOD 成功率提升到 0.9141，归一化存活时间、净旋转圈数和轴向平均角速度也同步提升。相较 Deploy-Base，其 OOD 线速度由 4.4935 cm/s 降至 4.2818 cm/s，平均力矩由 2.4433 降至 2.4087，表明成功率提升同时伴随更平稳的物体运动和更低的关节命令代价。若以 Teacher 作为 OOD 参考上界，Deploy-Refined 相比 Deploy-Base 在成功率、归一化存活时间、净旋转圈数与轴向平均角速度上的上界差距分别收缩约 38.46%、63.08%、50.12% 与 53.53%。因此，本文最终采用的第二阶段增强策略在常规圆柱体 OOD 评测中显著缩小了部署策略与教师上界之间的差距。

综合图 4.8、图 4.9 与表 4.3，第二阶段增强使可部署策略在圆柱体参数分布外条件下获得了更强的总体适应能力。为了进一步说明这种提升来自哪些具体设计，表 4.4 将第二阶段增强项拆分为质心分布放宽和动作一致性约束两部分进行对比。

表 4.4 第二阶段增强设计的 OOD 消融结果

策略设置	成功率 ↑	存活时间 ↑	圈数 ↑	角速度 ↑	动作差异 ↓
Deploy-Base	0.8750	0.9492	3.0422	0.9803	–
仅放宽质心分布	0.8750	0.9469	3.0312	0.9847	0.3929
仅动作一致性约束	0.8906	0.9640	3.2001	1.0314	0.3635
质心放宽 + 动作一致性	**0.9141**	**0.9758**	**3.2558**	**1.0420**	**0.3053**

表 4.4 表明，单独放宽质心偏移分布并未超过 Deploy-Base，说明仅增加高风险样本并不足以保证历史编码器学到更可用的部署表征；单独加入动作一致性约束后，OOD 成功率提高到 0.8906，净旋转圈数和轴向平均角速度也同步提升；两项设计联合使用时，成功率进一步达到 0.9141，动作差异降至 0.3053。表 4.9 进一步比较了 $\lambda_{\text{act}} \in \{0, 0.25, 0.5, 1.0, 2.0\}$ 的候选设置，其中 $\lambda_{\text{act}} = 1.0$ 在本组实验中取得最高 OOD 成功率并保持较低动作差距，因此本文将其作为最终增强策略的固定权重。总体来看，动作层面的教师——部署一致性是主要增益来源，更宽的质心分布则为联合训练提供额外的风险样本覆盖。

4.3 消融实验与设计选择分析

本小节汇总用于支撑关键设计选择的代表性消融实验。以下表格分别对应特权信息类别、技能先验中间表征、历史窗口长度、历史编码器结构和动作一致性权重；第二阶段增强项的组合消融已经在表 4.4 中给出，这里不再重复列出相同表格。

表 4.5 汇总了特权信息消融在最终部署策略效果上的影响。该组实验固定第二阶段部署形式，主要研究第一阶段教师可使用的信息类别改变后，最终部署策略是否仍能形成可迁移的旋转能力。结果表明，去除物体位置、质量与摩擦或质心偏移都会削弱 OOD 表现，其中缺失质心偏移时的部署策略在 ID 与 OOD 条件下均未达到成功判据，表明当前任务对偏心引起的接触变化更为敏感。需要注意的是，该表并非九维特权信息的逐维完备消融，而是保留了三类与任务机制直接相关的代表性对照，用于解释第四章中偏置圆柱体条件下的退化现象。

表 4.5 策略引入的特权信息消融结果

部署策略设置	ID	OOD				
	成功率↑	成功率↑	存活时间↑	圈数↑	角速度↑	线速度↓	力矩↓
完整特权信息	**0.9922**	**0.8750**	**0.9492**	**3.0422**	**0.9803**	**4.4935**	**2.4433**
去除物体位置	0.9453	0.6172	0.7419	2.3078	0.9231	5.0319	4.1555
去除质量与摩擦	0.9375	0.7344	0.8680	2.6264	0.9073	4.8292	2.5439
去除质心偏移	0.0000	0.0000	0.2306	0.4247	0.5707	4.7125	2.9787

表 4.6 给出技能先验中间表征消融结果。该对照核心是比较特权信息进入策略主干前是否经过任务相关的中间表征映射。直接使用 9 维特权输入后，教师 OOD 成功率、部署 ID 成功率和部署 OOD 成功率均下降，部署侧角速度明显降低，线速度反而升高，表明原始物理参数直接拼接并没有自动带来更好的可部署策略。相较之下，9D → 8D 技能先验在当前网络规模和训练预算下起到了表征约束作用，使第一阶段动作主干条件和第二阶段历史监督目标保持在同一较稳定的先验空间中。

表 4.6 技能先验中间表征消融结果

表征设置	ID	OOD			
	部署成功率↑	教师成功率↑	部署成功率↑	部署角速度↑	部署线速度↓
9D → 8D 技能先验	**0.9922**	**0.9766**	**0.8750**	**0.9803**	**4.4935**
9D 特权直接输入	0.7734	0.8672	0.5391	0.4005	5.0972

表 4.7 给出第二阶段历史长度消融结果。历史窗口的作用是让编码器从最近一段本体感觉响应中估计物体接触和动力学差异，因此过短可能缺少动态线索，过长则会增加输入冗余和部署缓存开销。结果显示，不同历史长度均能维持较高 ID 成功率，但 OOD 指标并不随窗口长度单调提高；25 帧在成功率、线速度和力矩上较好，30 帧则取得该组最高 OOD 旋转圈数和轴向角速度。本文因此采用 30 帧作为历史覆盖、旋转效率和部署开销之间的折中选择，并不意味着其为唯一最优的窗口长度，在各个方面都有最好的执行效果。

表 4.7 第二阶段历史长度消融结果

历史长度	ID	OOD				
	成功率↑	成功率↑	圈数↑	角速度↑	线速度↓	力矩↓
25	**1.0000**	**0.9219**	3.1900	1.0178	**4.1640**	**2.3925**
28	0.9688	0.8203	2.9729	0.9914	4.5477	2.4996
30	0.9922	0.8906	**3.2301**	**1.0500**	4.3687	2.4110
35	0.9922	0.8594	3.0883	1.0039	4.5418	2.4156
40	0.9766	0.9063	3.1772	1.0265	4.3824	2.4103
45	0.9922	0.8906	3.1374	1.0093	4.3431	2.4081

表 4.8 给出历史编码器结构消融结果。Flatten-MLP 将历史窗口直接展平，虽然实现简单，但较难显式利用相邻时刻之间的局部变化，因此其 OOD 成功率、先验误差和动作差距均落后于时序结构。1D-CNN 在先验误差、动作差距和 OOD 角速度上表现最好，表明局部时序模式已经能提供有效的适应线索；GRU 在本次单种子评测中取得更高成功率，但角速度和动作对齐指标未超过 1D-CNN。综合并行推理便利性、部署端无需维护递归隐状态以及旋转效率，本文采用一维时序卷积作为默认历史编码器，同时保留递归结构作为后续优化方向。

表 4.8 第二阶段编码器结构消融结果

编码器结构	ID	OOD				
	成功率↑	成功率↑	角速度↑	力矩↓	先验误差↓	动作差距↓
Flatten-MLP	0.9688	0.7969	0.9036	**2.3930**	0.2285	0.4597
1D-CNN	0.9922	0.8906	**1.0500**	2.4110	**0.1242**	**0.3018**
GRU	**1.0000**	**0.9141**	1.0176	2.3972	0.1273	0.3147

正文表 4.4 已经给出第二阶段增强项的组合消融，说明动作一致性约束与质心分布放宽联合使用时能够将常规 OOD 成功率提高到 0.9141。在此基础上，表 4.9 进一步给出动作一致性损失权重消融结果。该组实验均使用更宽质心分布的第二阶段训练设置，仅改变 λ_{act}，用于研究动作模仿项过弱或过强时对部署性能的影响。结果显示，较小权重虽然能够带来一定提升，但动作差距不一定降低；而当权重增大到 2.0 时，OOD 成功率反而下降，表明过强的一致性约束可能压缩历史编码器根据自身观测进行适应的空间。综合成功率、存活时间、旋转效率和动作差距，$\lambda_{act} = 1.0$ 在本组对比中均最优，因此本文采用该权重作为最终实验设置。

表 4.9 动作一致性损失权重消融结果

λ_{act}	ID	OOD				
	成功率↑	成功率↑	存活时间↑	圈数↑	角速度↑	动作差距↓
0	0.9766	0.8750	0.9469	3.0312	0.9847	0.3929
0.25	**1.0000**	0.8828	0.9551	3.1595	1.0099	0.4317
0.5	0.9922	0.8906	0.9417	3.1117	1.0092	0.4594
1.0	**1.0000**	**0.9141**	**0.9758**	**3.2558**	**1.0420**	**0.3053**
2.0	**1.0000**	0.8672	0.9483	3.0868	1.0049	0.3466

4.4 实机部署与评估

完成仿真评测后，本文将同一个 Deploy-Refined checkpoint 直接迁移到真实 LEAP Hand。实机端不提供物体质量、摩擦、质心等训练期特权信息，策略只能读取关节位置、关节目标及其历史序列；底层仍按照第二章定义的相对关节目标增量执行位置控制。这里更关心两个实际问题：策略离开仿真后能否闭环运行，以及圆柱体和少量近圆柱对象会在哪些条件下失稳。全部实机测试均不针对单个物体重新训练，也不单独调整策略参数。

为避免启动瞬间破坏接触，系统先用 5 s warmup 缓慢移动到初始抓取姿态，并在初始抓取中采用 0.10 的闭合系数。正式控制时，循环频率保持 30 Hz，动作尺度仍为 $1/24$，电机 P/D 增益分别设为 600 和 150，单电机电流限制为 400 mA，串口波特率为 4,000,000。程序加载历史编码器、冻结动作主干和训练阶段保存的归一化统计量后，用当前关节状态初始化短历史与 30 帧长历史缓存；随后每个控制周期更新本体历史、估计技能先验、输出相对关节目标增量，并在动作截断和关节限位后发送给电机。归一化统计量、历史填充方式、动作缩放和关节限位必须与训练阶段保持同一语义，否则仿真中已经收敛的策略在真实平台上也可能表现为动作幅值异常或接触迅速破坏。

实机评测协议单独定义，避免与仿真统一评测指标混用。仿真成功率要求回合未掉落且净旋转角不少于一圈；实机每类对象进行 10 次试验、每次最长 20 s，若物体未因掉落、明显倾倒或卡棱提前失效，并产生可辨识的受控旋转，则记为成功。实机净旋转圈数由顶部标记或瓶身标签的方向变化人工估计，轴向角速度再由估计圈数和测试时长换算得到。因此，这里的圈数和角速度用于描述真实平台上的低速执行现象，不作为外部姿态传感器意义上的精确测量。

图 4.10 列出了四个实机测试对象。对象按照难度递增排列：训练分布附近的标准圆柱体、尺寸外推的圆柱体、细高饮料瓶，以及带棱角的饮料瓶。前两个对象主要改变圆柱

尺寸、质量和质心，第三个对象在近圆截面基础上增加长径比和瓶颈结构，第四个对象进一步引入纵向筋线和局部凹凸。这样的对象顺序把“圆柱参数变化”“近圆柱迁移”和“非圆截面接触”分开，便于观察策略从哪一类几何变化开始失效。

(a) 标准圆柱体　(b) 尺寸外圆柱体　(c) 细高饮料瓶　(d) 带棱角饮料瓶

图 4.10　实机部署测试对象集

对象属性与定量结果如表 4.10 所示。

表 4.10　实机部署对象物理属性与评估结果

对象	物理属性	成功次数	净旋转圈数	轴向角速度	失效模式
标准圆柱体	PLA，直径 7.5 cm，高 7.5 cm，90 g	10/10	约 0.5 圈	0.157 rad/s	无明显失效
尺寸外圆柱体	PLA，直径 5.5 cm，高 7.5 cm，55 g	10/10	约 1.0 圈	0.314 rad/s	无明显失效
细高饮料瓶	PET，直径 5.8 cm，高 18.0 cm，30 g	8/10	约 1.0 圈	0.314 rad/s	少数次倾倒掉落
带棱角饮料瓶	PET，直径 6.6 cm，高 22.0 cm，38 g，带纵向棱角	0/10	未形成稳定净旋转	–	卡棱导致推进中断

注：实机成功判据为 20 s 内未掉落、未明显倾倒或卡滞，并产生可辨识的受控旋转；该定义不同于仿真中的一圈净旋转成功标准。净旋转圈数和轴向角速度由视频标记人工估计得到。

表 4.10 显示，最终部署策略在两个 3D 打印圆柱体对象上均达到 100% 的实机成功率，说明仿真中学到的圆柱体旋转技能可以转化为真实 LEAP Hand 上可重复的闭环行为。不过，真实旋转速度明显低于仿真：标准圆柱体在 20 s 内约转 0.5 圈，尺寸外圆柱体约转 1.0 圈。仿真和实机的名义控制频率都约为 30 Hz，所以速度下降不能简单归因于频率不同。更直接的限制来自真实执行链路：电机受到 400 mA 电流上限和保守 P/D 增益约束，关节跟踪存在滞后；3D 打印圆柱与指尖之间的摩擦、柔顺性和局部接触面积也难以与仿真完全一致；串口通信、关节背隙和接触微滑还会削弱短时本体历史中的动作——状态对应关系。因而，该角速度更适合作为真实平台低速闭环执行的观察量，而不宜与仿真

旋转效率直接等价比较。

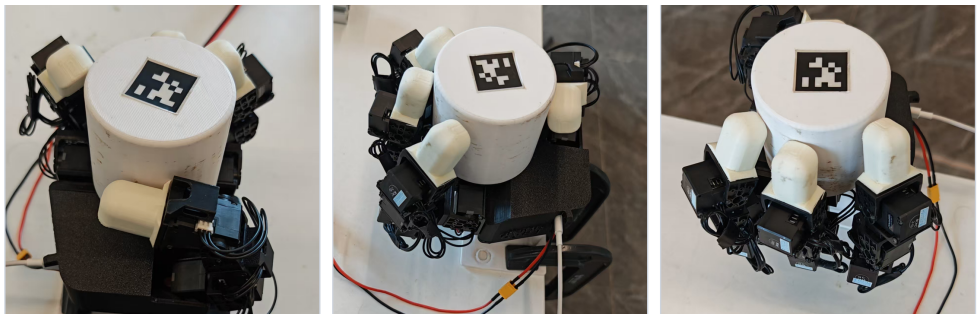

图 4.11 标准圆柱体实机旋转关键帧

图 4.11 给出了标准圆柱体的实机关键帧。圆柱体顶部标记在不同帧之间发生了可辨识的方向变化，说明策略确实推动物体绕竖直轴旋转；四指在整个过程中仍保持包络，没有出现明显张开、滑脱或被挤出手内区域的现象。对于几何尺寸接近训练对象、表面为硬质圆柱面的样本，本体历史适应策略可以在真实平台上维持低速连续旋转。

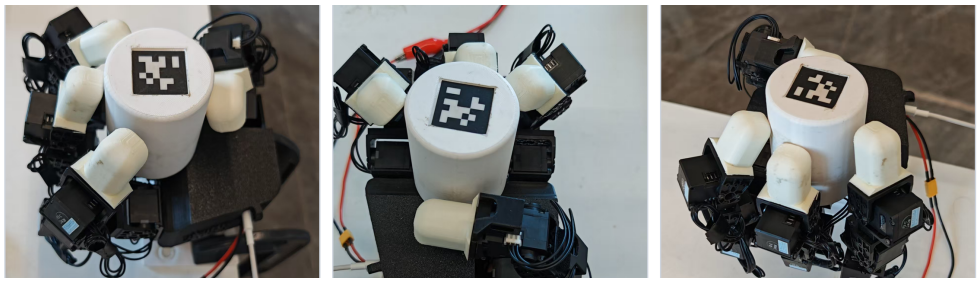

图 4.12 尺寸外圆柱体实机旋转关键帧

图 4.12 展示了尺寸外圆柱体的执行过程。该对象直径更小，手指闭合后的接触半径随之改变，对接触位置和关节目标跟踪提出了额外要求。关键帧中，顶部标记产生了比标准圆柱体更大的方向变化，物体也基本保持在手掌中心附近，没有明显掉落。这与表 4.10 中约 1.0 圈的净旋转估计相一致，说明当前策略对圆柱尺寸变化具有一定适应能力。

细高饮料瓶进一步引入了更高的长径比、PET 材料和瓶颈结构。图 4.13 中，瓶身标签方向发生变化，说明策略在近圆柱真实物品上仍能产生有效旋转趋势，并在多数试验中维持到 20 s 上限。相比两个 3D 打印圆柱体，细高瓶更容易出现上端摆动和整体倾斜，少量试验会因倾斜积累而掉落。因此，80% 的成功率只能说明策略具备初步近圆柱实物迁移能力，稳定性仍弱于规则圆柱体。

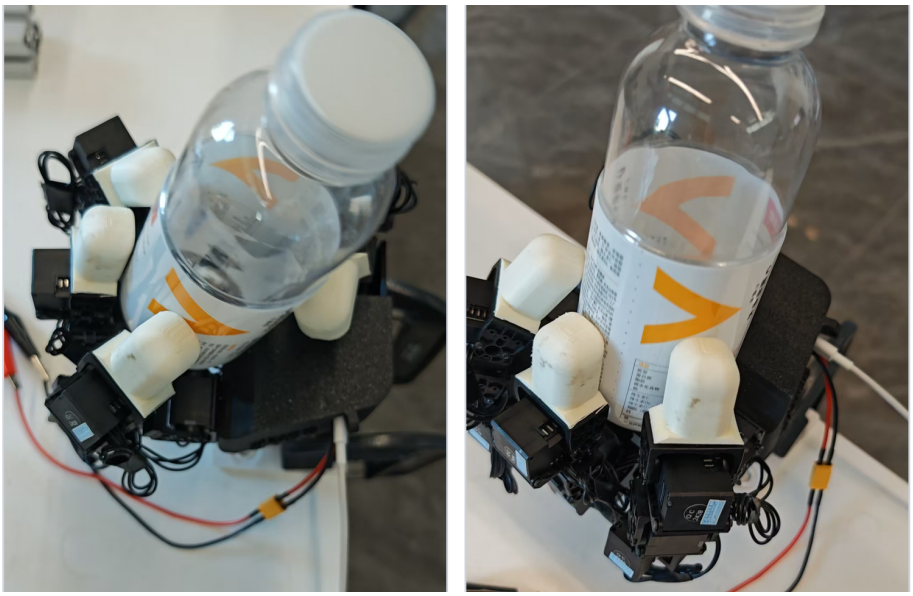

图 4.13 细高饮料瓶实机旋转关键帧

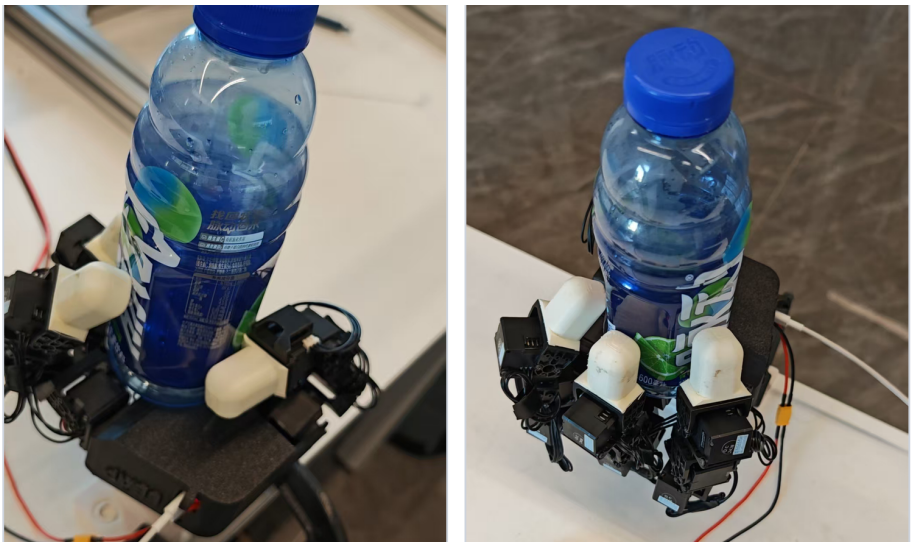

图 4.14 带棱角饮料瓶实机卡棱关键帧

带棱角饮料瓶暴露了当前方法的主要局限。图 4.14 中，手指能够完成初始抓持，也能对瓶身产生局部推动；但瓶身纵向筋线和凹凸结构会使接触点在旋转过程中发生突变，指尖容易从连续推动转变为抵住棱边。物体虽然有轻微旋转趋势，却难以形成持续的净位姿推进，10 次试验均未计为成功。也就是说，当前策略主要掌握的是圆柱或近圆柱表面上的连续接触迁移规律，对非圆截面和局部棱边结构的几何泛化仍然不足。

综合来看，同一 Deploy-Refined 策略可以在不引入真实物体特权参数、不针对单个对象重新训练的条件下，在真实 LEAP Hand 上闭环运行。标准圆柱体和尺寸外圆柱体的 100% 成功率说明，仿真中学到的圆柱体连续旋转技能能够迁移到真实硬件，并形成可重

复的受控旋转行为；细高饮料瓶的结果进一步显示出一定近圆柱迁移能力。带棱角饮料瓶上的卡棱失败则划出了当前方法的边界：策略尚不能稳定处理非圆截面和局部棱边导致的接触突变，因此还不能被解释为任意对象的稳定泛化旋转。

4.5 本章小结

本章围绕 LEAP Hand 圆柱体连续手内旋转任务给出了实验设置、仿真结果与实机部署评估。第一阶段教师策略已经学到稳定的连续旋转技能；第二阶段原始可部署链路能够在仅依赖本体感觉历史的条件下完成连续旋转，而以教师——部署动作一致性为主、并辅以质心偏移分布放宽的增强策略，则把常规圆柱体 OOD 成功率提升到 0.9141，缓解了第二阶段退化。实机侧，同一 Deploy-Refined 策略经过观测、历史、动作和电机控制侧的 Sim2Real 对齐后，能够在真实 LEAP Hand 上完成闭环部署，并在标准圆柱体、尺寸外圆柱体以及部分近圆柱真实对象上形成连续、可重复的受控旋转。真实速度低于仿真以及带棱角饮料瓶上的卡滞也说明，执行器动力学、接触物理和对象局部几何仍是后续提升真实泛化能力的关键限制。本文已经验证圆柱体连续旋转、圆柱体参数泛化与少量真实对象部署的可行性，更大范围的多物体稳定泛化仍留作后续工作。

5 结论与展望

5.1 工作总结

本文围绕 LEAP Hand 圆柱体连续手内旋转任务展开研究，目标是在低成本灵巧手平台上建立从仿真建模、策略训练到真实部署的完整链路。全文工作可概括为三部分。

在任务定义层面，本文将连续手内旋转落到可评测、可训练的具体场景中。围绕圆柱体对象和竖直轴连续旋转目标，论文明确了系统状态、策略观测、训练期特权信息、相对关节目标动作空间、轴向旋转量计算方式、奖励函数和回合终止条件。这些定义既保留连续旋转任务对旋转效率和抓持稳定性的要求，也使策略输入与真实 LEAP Hand 可读取的本体感觉信息保持一致。

在策略设计层面，本文构建了技能先验驱动的本体自适应旋转框架。第一阶段利用仿真中的特权信息学习旋转技能先验和教师策略，使系统在对象属性已知的训练条件下形成可持续旋转能力；第二阶段冻结动作主干，训练历史编码器从本体感觉历史中恢复同一技能先验，从而得到可部署策略；针对第二阶段在分布外条件下的退化，本文又引入教师——动作一致性约束，并配合更宽的质心偏移训练分布，对部署模块进行定向补强。

在实验验证层面，本文完成了仿真评测和实机部署两部分闭环。仿真中，教师策略能够稳定完成连续旋转，增强后的部署策略在常规分布外评测中取得 0.9141 的成功率；实机中，同一策略能够在真实 LEAP Hand 上闭环运行，并在标准圆柱体、尺寸变化圆柱体和部分近圆柱真实物体上实现低速受控旋转。与此同时，带棱角饮料瓶的失败也说明，当前方法对“近圆柱”对象仍然更稳，对明显非圆柱对象的泛化还不够。

5.2 研究结论

结合第四章的仿真和实机结果，本文得到以下结论。

第一，圆柱体连续手内旋转可以在低成本 LEAP Hand 平台上通过仿真强化学习获得可用能力。第一阶段教师策略在训练后期能够保持较高的轴向角速度、较低的掉落率和接近完整回合的执行时长，说明当前任务建模和奖励设计能够引导策略学到稳定旋转，而不是只学到短时摆动。

第二，本体历史自适应能够将教师侧能力转化为可部署策略，但对分布外参数变化的耐受度有限。原始第二阶段策略在 ID 条件下已经接近教师表现；当对象质量分布和质心偏移范围向外扩展后，成功率、净旋转圈数和轴向角速度均下降。仅靠本体历史恢复技能先验，尚不足以完全覆盖高风险物理参数带来的偏移。

第三，教师——动作一致性约束是本工作中最有效的增强手段。该约束不改变策略输入形式，而是直接约束部署侧输出贴近教师侧动作语义，因此在 OOD 条件下带来更直接的收益。放宽质心偏移分布也有作用，但它主要补充高风险样本覆盖，并不能单独解决泛化问题。

第四，仿真到真实的差异主要来自执行器、接触条件和观测稳定性，而不是任务目标本身。真实平台的速度上限、通信延迟、关节背隙和接触微滑都会压低实际旋转效率，因此实机结果不应与仿真数值一一对应。即便如此，同一部署策略仍能在真实圆柱体和部分近圆柱物体上形成可重复的闭环旋转，说明本文方法链路具备实际可执行性。

5.3 后续展望

围绕本文已经暴露出的边界，后续工作可从三个方向继续推进。

第一，可将真实执行器特性更细致地纳入仿真。例如，在训练阶段加入电机电流限制、通信延迟、关节背隙和更接近实机的动力学模型，使策略在仿真中提前面对更真实的执行条件，从而降低后续实机迁移落差。

第二，可将实机评测从可执行性判断推进到稳定性度量。目前旋转圈数和角速度主要依赖视觉标记或人工观察，适合做结果判断，但不足以支撑更细的控制分析。后续若加入外部视觉追踪、轻量级姿态标记或多视角估计，实机与仿真的对比会更完整。

第三，可继续扩展对象范围，但这一方向需要更谨慎。带棱角饮料瓶的失败说明，当前策略主要掌握的是圆柱或近圆柱表面的连续接触迁移规律。若要进一步走向多物体泛化，需要同时处理真实动力学建模、对象几何覆盖、姿态测量精度和高风险参数采样等问题，不能仅依赖增加训练轮数。

本文给出了低成本灵巧手实现连续手内旋转的一条可行路径，而更复杂对象上的稳定泛化，仍需在后续工作中继续验证和推进。

参考文献

[1] BICCHI A. Hands for dexterous manipulation and robust grasping: a difficult road toward simplicity[J/OL]. IEEE Transactions on Robotics and Automation, 2000, 16(6): 652-662[2026-05-03]. https://doi.org/10.1109/70.897777.

[2] OKAMURA A M, SMABY N, CUTKOSKY M R. An overview of dexterous manipulation[C/OL]//Proceedings 2000 IEEE International Conference on Robotics and Automation. San Francisco, CA, USA: IEEE, 2000: 255-262[2026-05-03]. https://doi.org/10.1109/ROBOT.2000.844067.

[3] OPENAI, ANDRYCHOWICZ M, BAKER B, et al. Learning dexterous in-hand manipulation[J/OL]. The International Journal of Robotics Research, 2020, 39(1): 3-20[2026-05-03]. https://doi.org/10.1177/0278364919887447.

[4] XIA Z, DENG Z, FANG B, et al. A review on sensory perception for dexterous robotic manipulation[J/OL]. International Journal of Advanced Robotic Systems, 2022, 19(2): 17298806221095974[2026-05-03]. https://doi.org/10.1177/17298806221095974.

[5] HOGAN N. Impedance control: an approach to manipulation, part I: theory[J/OL]. Journal of Dynamic Systems, Measurement, and Control, 1985, 107(1): 1-7[2026-05-03]. https://doi.org/10.1115/1.3140702.

[6] TOBIN J, FONG R, RAY A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C/OL]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, BC, Canada: IEEE, 2017: 23-30[2026-05-03]. https://doi.org/10.1109/IROS.2017.8202133.

[7] PENG X B, ANDRYCHOWICZ M, ZAREMBA W, et al. Sim-to-real transfer of robotic control with dynamics randomization[C/OL]//2018 IEEE International Conference on Robotics and Automation. Brisbane, QLD, Australia: IEEE, 2018: 3803-3810[2026-05-03]. https://doi.org/10.1109/ICRA.2018.8460528.

[8] OPENAI, AKKAYA I, ANDRYCHOWICZ M, et al. Solving Rubik's cube with a robot hand[EB/OL]. arXiv:1910.07113, 2019[2026-05-03]. https://arxiv.org/abs/1910.07113.

[9] MAKOVIYCHUK V, WAWRZYNIAK L, GUO Y, et al. Isaac Gym: high performance GPU-based physics simulation for robot learning[EB/OL]. arXiv:2108.10470, 2021[2026-05-03]. https://arxiv.org/abs/2108.10470.

[10] SHAW K, AGARWAL A, PATHAK D. LEAP Hand: low-cost, efficient, and anthropomorphic hand for robot learning[C/OL]//Robotics: Science and Systems. Daegu, Republic of Korea, 2023[2026-05-03]. https://www.roboticsproceedings.org/rss19/p089.pdf.

[11] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. arXiv:1707.06347, 2017[2026-05-03]. https://arxiv.org/abs/1707.06347.

[12] RAJESWARAN A, KUMAR V, GUPTA A, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations[C/OL]//Robotics: Science and Systems. 2018[2026-05-03]. https://doi.org/10.15607/RSS.2018.XIV.049.

[13] ZHU H, GUPTA A, RAJESWARAN A, et al. Dexterous manipulation with deep reinforcement learning: efficient, general, and low-cost[C/OL]//2019 International Conference on Robotics and Automation. Montreal, QC, Canada: IEEE, 2019: 3651-3657[2026-05-03]. https://arxiv.org/abs/1810.

06045.

[14] VAN HOOF H, HERMANS T, NEUMANN G, et al. Learning robot in-hand manipulation with tactile features[C/OL]//2015 IEEE-RAS 15th International Conference on Humanoid Robots. Seoul, Republic of Korea: IEEE, 2015: 121-127[2026-05-03]. https://doi.org/10.1109/HUMANOIDS.2015.7363524.

[15] QI H, KUMAR A, CALANDRA R, et al. In-hand object rotation via rapid motor adaptation[C/OL]//Proceedings of The 6th Conference on Robot Learning. PMLR, 2023, 205: 1722-1732[2026-05-03]. https://proceedings.mlr.press/v205/qi23a.html.

[16] QI H, YI B, SURESH S, et al. General in-hand object rotation with vision and touch[C/OL]//Proceedings of The 7th Conference on Robot Learning. PMLR, 2023, 229: 2549-2564[2026-05-03]. https://proceedings.mlr.press/v229/qi23a.html.

[17] YANG M, LU C, CHURCH A, et al. AnyRotate: gravity-invariant in-hand object rotation with sim-to-real touch[C/OL]//Proceedings of The 8th Conference on Robot Learning. PMLR, 2025, 270: 4727-4747[2026-05-03]. https://proceedings.mlr.press/v270/yang25c.html.

[18] YUAN H, ZHOU B, FU Y, et al. Cross-embodiment dexterous grasping with reinforcement learning[C/OL]//International Conference on Learning Representations. 2025[2026-05-03]. https://openreview.net/forum?id=twIPSx9qHn.

[19] KUMAR A, FU Z, PATHAK D, et al. RMA: rapid motor adaptation for legged robots[C/OL]//Robotics: Science and Systems. 2021[2026-05-03]. https://doi.org/10.15607/RSS.2021.XVII.011.

[20] MITTAL M, YU C, YU Q, et al. Orbit: a unified simulation framework for interactive robot learning environments[J/OL]. IEEE Robotics and Automation Letters, 2023, 8(6): 3740-3747[2026-05-03]. https://doi.org/10.1109/LRA.2023.3270034.