Jingfeng

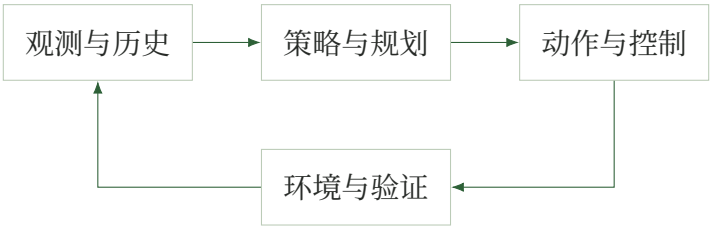
具身智能

从感知与行动闭环到机器人基础模型

Jingfeng · 2026 年 10 月 22 日修订



让每一次预测回到物理反馈中



本书从可执行的机器人接口出发，串起感知、行为克隆、生成式动作、机器人基础模型、规划与实验。公式说明假设，算法说明输入输出，评测追踪干预与失败。

原稿：2026 年 9 月讲义（原资料截点 2026-09-03）。本版完成全文修订、独立教学实验与 Jingfeng 排版迁移；证据核验日期为 2026-10-22。模型介绍使用明确的论文版本，不声称穷尽该日期之前的所有工作。

如何使用这本书

这本书的主线是一条因果链：机器人获得不完整观测，以历史或信念表示当前信息，选择动作，经控制器改变世界，再用新的观测验证结果。每种网络架构、损失和数据配方，都放到这条链中解释。

读者需要线性代数、基础概率和梯度下降知识，能阅读 Python 代码即可开始。Transformer、相机几何、机器人运动学和强化学习会在使用时交代必要定义；涉及更深入的潜变量学习、通用强化学习理论和量化内核时，分别指向本系列《世界模型》《强化学习》《智能体模型量化压缩与部署》的真实主题，不预设其他分册的页码。

三条阅读路线

建立系统观 先读第一 – 三章，再读第八章和第十六章；手算第二章的二维接口与延迟例子。

复现动作学习 读第四 – 五章、第九章和第十三章，再运行第十八章的轻量实验；最后对照第六章拆解一个公开模型。

开展研究 在数据、基准、安全三章明确研究边界，再从可选研究练习中挑选一个假设；用消融、负结果与日志缩小问题。

如何理解书中的证据

- **定义与推导**：在正文给出假设和符号，不能把最坏界、经验规律和形式保证互相替换。
- **论文报告**：模型规模、数据与机制以本版参考文献的版本为准，核验日期为 2026-10-22；介绍一篇预印本不等于独立复现其结论。无法可靠核对的细粒度性能数字不用于排名。
- **手算与合成教学数据**：二维两连杆示例、人工编写的评测日志用于验证接口和统计口径，不代表真实机器人实验。
- **本次实测**：配套 NumPy 教学仿真的行为克隆消融有独立配置、数据划分、种子、逐回合结果与命令。它证明的是该环境下的观测与时域效应，不能外推为 LIBERO、机械臂或真实世界成绩。

本书不以方法名称数量代替解释。阅读每篇论文时，先回答：观测是什么，动作是什么，策略看多长历史，控制周期多长，数据如何划分，损失监督什么，成功怎样定义，失败时谁接管。没有这些答案，单个成功率无法说明系统能力。

复现顺序

先读动作 schema，再检查一个样本的反归一化和坐标变换；通过小数据过拟合后运行闭环；最后才扩大模型。练习旁的提示帮助定位思路，附录统一提供公式和接口检查表。新版与归档稿独立存放，构建方法见本书 README。

目录

如何使用这本书 ii

第一部分　共同语言：从接口进入闭环 1

第一章　具身智能究竟是什么 2
- 1.1 从“识别世界”到“改变世界” . . . 2
- 1.2 具身不等于“有身体” . . . 2
- 1.3 三种尺度：技能、策略与系统 . . . 2
- 1.4 形态决定计算问题 . . . 3
- 1.5 任务先于模型：一个可复述的研究对象 . . . 3
- 1.6 一张领域地图 . . . 3

第二章　数学底座：信息、轨迹与动作接口 4
- 2.1 状态、观测、历史与信念 . . . 4
- 2.2 轨迹分布与回报的下标 . . . 4
- 2.3 价值与优势属于哪种信息状态 . . . 5
- 2.4 动作必须带坐标、单位和控制语义 . . . 5
- 2.5 二维两连杆到达：同一目标的两种描述 . . . 6
- 2.6 预测时域、执行时域与延迟手算 . . . 7
- 2.7 目标与约束是不同的接口 . . . 7

第三章　机器人系统栈：端到端模型之外还有什么 8
- 3.1 经典模块化栈 . . . 8
- 3.2 现代学习系统的三种集成方式 . . . 8
- 3.3 高层慢思考与低层快控制 . . . 8
- 3.4 本体状态为何不能省 . . . 9
- 3.5 标定怎样变成模型可用的信息 . . . 9
- 3.6 跨模态同步和策略输入 . . . 9
- 3.7 眞正的系统接口清单 . . . 10

第二部分　表示与决策：从动作学习到长程执行 11

第四章　多模态感知与表示 12
- 4.1 视觉表示：语义与空间精度的张力 . . . 12
- 4.2 二维、三维与多视角 . . . 12
- 4.3 语言条件究竟提供什么 . . . 12
- 4.4 触觉、力与音频 . . . 13
- 4.5 从模态编码到动作条件 . . . 13
- 4.6 token 压缩与信息瓶颈 . . . 13

第五章　动作生成：回归、token、扩散与流匹配 14
- 5.1 回归为何会落在两个有效动作之间 . . . 14

5.2 动作 token：量化、训练与解码 . . . 14
5.3 Chunk 不是一种独立损失 . . . 15
5.4 Diffusion Policy：训练一个去噪器 . . . 15
5.5 Flow matching：学习从噪声走向动作的速度 . . . 16
5.6 按失效条件选择动作头 . . . 16

第六章 VLA 与机器人基础模型的架构谱系 18
6.1 先比较接口，再比较主干 . . . 18
6.2 从机器人专用策略到网页语义共训 . . . 18
6.3 开放策略与连续动作专家 . . . 19
6.4 动作专家、双系统与小模型的其他选择 . . . 21
6.5 从模型卡选出基线 . . . 22

第七章 世界模型、规划与想象 23
7.1 机器人规划需要什么后果预测 . . . 23
7.2 MPC：规划多步，只执行有限步 . . . 23
7.3 视频预测、动作 grounding 与逆动力学 . . . 23
7.4 把组合系统变成可检查的循环 . . . 24

第八章 长程具身 Agent：计划、技能、记忆与恢复 25
8.1 把长任务拆成可验证的有限执行 . . . 25
8.2 SayCan：语言相关性与当前可执行性 . . . 25
8.3 一个可审计的技能契约 . . . 25
8.4 记忆要能更新，也要允许不知道 . . . 26
8.5 验证、恢复与评测必须共用同一条轨迹 . . . 26

第三部分 训练与部署：从示范到交互 27

第九章 行为克隆、分布偏移与交互式模仿 28
9.1 行为克隆：在历史上做条件预测 . . . 28
9.2 误差为何复合：上界及其假设 . . . 28
9.3 DAgger：让专家标注学习者会访问的历史 . . . 29
9.4 GAIL：匹配哪些占用分布 . . . 29
9.5 序列 BC、数据增强与重标注 . . . 30
9.6 从离线诊断走向闭环实验 . . . 30

第十章 从离线数据到强化学习后训练 31
10.1 先定义决策单位与信息条件 . . . 31
10.2 最大熵目标与 SAC . . . 31
10.3 离线 RL：策略为何会利用数据外高估 . . . 32
10.4 机器人后训练与语言模型后训练的关系 . . . 32
10.5 奖励、成本与接管日志 . . . 33
10.6 扩散与 flow 策略怎样接受后果反馈 . . . 33

第十一章 数据混合、适配与跨本体学习 34
11.1 预训练与后训练分别解决什么 . . . 34
11.2 采样权重与损失权重不要混为一谈 . . . 34

11.3 共训练：哪类信息能传到控制 . . . 34
11.4 冻结、LoRA 与全量微调的选择 . . . 35
11.5 跨本体对齐：共享数组之前先共享含义 . . . 35
11.6 训练稳定性与验证顺序 . . . 36

第十二章 Sim2Real：辨识差距，再选择适配方法 37
12.1 域随机化的目标与覆盖条件 . . . 37
12.2 系统辨识先缩小不确定性 . . . 37
12.3 残差学习：保留基控制器的条件 . . . 37
12.4 数字孪生与合成示范 . . . 38
12.5 把转移失败分解成最小诊断 . . . 39

第十三章 训练与部署中的时序、归一化和执行 40
13.1 从曝光到执行的时间戳 . . . 40
13.2 动作归一化：统计、逆变换与饱和 . . . 40
13.3 预测时域、执行时域和异步队列 . . . 40
13.4 自适应执行长度是一项待验证策略 . . . 41
13.5 缓存、编译和压缩如何影响闭环 . . . 41
13.6 终止、恢复与人工接管 . . . 41
13.7 一个可审查的训练循环接口 . . . 42

第四部分 数据与证据：让闭环表现可以检验 43

第十四章 数据工程、schema 与无泄漏划分 44
14.1 数据规模先还原到物理时间 . . . 44
14.2 多源数据的对齐强度与用途 . . . 44
14.3 最小 schema：结构可检查，物理语义可追踪 . . . 44
14.4 代表资源：按用途而非单一规模排列 . . . 47
14.5 按主张设计 group split . . . 48
14.6 质量、失败数据与发布责任 . . . 48

第十五章 仿真工具与不可混同的任务协议 49
15.1 按科学问题选择工具 . . . 49
15.2 导航与家庭交互提供什么证据 . . . 49
15.3 RLBench：task、variation 和 episode 三层 . . . 50
15.4 CALVIN：连续子任务链与环境泛化 . . . 50
15.5 LIBERO：原始终身学习与后续多任务协议 . . . 50
15.6 为什么三个基准的百分比不能横向排名 . . . 51
15.7 可复现实验配置应落到文件 . . . 51

第十六章 闭环评测、统计区间与失败恢复 52
16.1 成功率的 Wilson 区间 . . . 52
16.2 六轴泛化：留出什么，固定什么 . . . 53
16.3 同一流程中的任务、时延、干预和安全指标 . . . 53
16.4 二十条合成记录怎样汇总 . . . 54
16.5 为什么离线动作误差不够 . . . 54
16.6 公平比较、故障与失败分析 . . . 55

第十七章 运行时安全、控制约束与保证边界 **56**
17.1 风险来自整个感知与执行链 . . . 56
17.2 分层机制与各自的证据 . . . 56
17.3 不确定性要经过风险校准 . . . 56
17.4 一个可推导的 CBF 条件 . . . 57
17.5 采样、延迟和估计误差会改变保证 . . . 58
17.6 形式化接口与可测试边界 . . . 58
17.7 故障注入与安全评测 . . . 58

第十八章 最小实验：动作时域、扰动与评测口径 **59**
18.1 本次运行：行为克隆与动作块长度 . . . 59
18.2 合成记录：接管、恢复与成功率口径 . . . 60
18.3 复现入口与边界 . . . 60
18.4 从最小结果走向机器人复现 . . . 60

第五部分 研究实践：把问题变成可证伪实验 **61**

第十九章 研究地图与前沿证据的边界 **62**
19.1 如何使用这张地图 . . . 62
19.2 从单步预测扩展到状态、计划与验证 . . . 62
19.3 从成功示范扩展到异质经验 . . . 62
19.4 动作表示和部署预算一起设计 . . . 62
19.5 近期材料：机制与仍待验证的部分 . . . 63

第二十章 仍未解决的核心科学问题 **64**
20.1 跨本体共享什么，执行时还要补什么 . . . 64
20.2 长尾覆盖与物理反事实 . . . 64
20.3 长程信用、可更新记忆与持续学习 . . . 64
20.4 接触与灵巧操作中的充分信息 . . . 64
20.5 拒绝、求助与风险覆盖曲线 . . . 65
20.6 效率误差何时会变成轨迹误差 . . . 65

第二十一章 如何开展一项可信的具身研究 **66**
21.1 先区分观测事实与待检验假设 . . . 66
21.2 让证据范围匹配主张 . . . 66
21.3 同时控制数值精度与时间 . . . 66
21.4 用最小实验建立独立参考 . . . 66
21.5 失败记录、版本与阶段性部署 . . . 67

第二十二章 可选研究练习：从机制假设到反证 **68**
22.1 练习一：混合精度收益来自数值还是时序 . . . 68
22.2 练习二：动作码本保留了轨迹结构吗 . . . 68
22.3 练习三：低秩残差能否修正 flow 轨迹 . . . 69
22.4 练习四：动态执行长度是否优于固定 chunk . . . 69
22.5 练习五：恢复能力与失败检测分开学习 . . . 69
22.6 练习六：压缩预测器是否保留动作排序 . . . 69

第二十三章　学习与复现路线：以交付物检查进度 **71**
23.1 第 1–2 周：接口、概率与时间线 . . . 71
23.2 第 3–4 周：建立可复现的行为克隆基线 . . . 71
23.3 第 5–6 周：动作表示与模型卡 . . . 71
23.4 第 7–8 周：数据与评测契约 . . . 71
23.5 第 9–10 周：验证、恢复与计算预算 . . . 71
23.6 第 11–12 周：一个机制与一份研究包 . . . 72

第六部分　附录：公式、接口与阅读工具 **73**

附录 A　核心公式与使用条件 **74**
A.1 行为克隆、chunk 与优势加权 . . . 74
A.2 SAC：完整状态假设、终止与采样分布 . . . 74
A.3 Diffusion 与 flow matching 的端点 . . . 74
A.4 MPC 与终端条件 . . . 75
A.5 成功率区间与分母 . . . 75

附录 B　数据、动作与运行接口检查表 **76**
B.1 数据接入前 . . . 76
B.2 闭环执行前 . . . 76
B.3 结果归档前 . . . 76

附录 C　论文阅读与证据记录模板 **77**

附录 D　术语与符号表 **78**

结语 **80**

参考文献 **81**

插图

1.1 当前具身智能的主要模块与数据闭环。强系统通常是组合，而非单一网络。. . . . 3

3.1 经典机器人闭环。学习模型可以替换某一模块，也可以跨多个模块端到端训练。. 8

14.1 多源数据通过表示与监督接口连接；每一类都需要独立的质量和覆盖证据。. . . . 44

17.1 安全机制需跨层配合；每一层都可能受模型、传感器、时序或硬件故障限制。. . 56

表格

1.1 具身研究中经常被混为一谈的三个尺度 . . . 3

2.1 动作接口示例。频率均为本书教学配置，不是型号规格。 6
2.2 一次异步预测的教学时间线；动作索引绑定目标执行时刻。 7

4.1 机器人视觉表示的主要选择 12

5.1 动作头的条件与诊断。多峰表达能力不是闭环优越性的排名。 17

6.1 先冻结问题，再比较设计轴；不构成成功率排名。 22

8.1 技能接口应返回可检查字段，具体阈值由平台标定。 25

10.1 把后训练配方落到机器人前必须回答的问题 32

11.1 适配策略及其需要验证的条件 35

12.1 Sim2Real 差距与可区分的诊断；先验证接口，再判断模型不足 39

14.1 数据与任务资源的接口重点；规模和可用字段应以所用 release 为准 47

15.1 常见仿真生态的用途与需要核查的条件 49
15.2 协议差异决定指标含义；表中为需具体化的接口，不是统一排名 51

16.1 建议同时报告的指标及分母 53
16.2 合成教学日志的口径演示，不能解释为实测模型性能 54

17.1 三类安全表述的对象和限度 57

18.1 二维教学仿真的本次实测。成功率为三个训练种子的均值；MSE 单位为 $(m/s^2)^2$。 59

19.1 随书近期材料的选择性索引，均为论文自报，未由本次实验复现。 63

22.1 按已有条件选择起步范围，资源需求不构成创新性评级。 70

D.2 核心符号，局部算法的额外符号在正文定义。 79

第一部分

共同语言：从接口进入闭环

第一章　具身智能究竟是什么

1.1 从“识别世界”到“改变世界”

普通视觉模型学习 $p(y \mid x)$，语言模型学习 token 条件分布；具身智能体面对的是一个会被自身动作改变、且无法一次看全的世界。一次决策不是独立样本，而是轨迹的一部分：

$$\tau = (s_0, o_0, a_0, r_0, \dots, s_{T-1}, o_{T-1}, a_{T-1}, r_{T-1}, s_T, o_T).$$

动作 a_t 改变下一状态 s_{t+1}，策略又只能通过有噪声观测 o_{t+1} 看见后果。这带来三种根本差异：

1. **数据非独立同分布**：模型的错误会把自己带到训练集没有覆盖的状态；
2. **目标延迟且稀疏**：早期抓取姿态可能在几十秒后才导致装配失败；
3. **错误具有物理代价**：碰撞、摔落、夹伤与硬件磨损不是“答错一道题”。

定义

本书的讨论对象是机器人与物理仿真中的**具身智能**：智能体通过有身体约束的传感与行动闭环完成任务，行动会改变未来状态和可获得的信息。跨任务、场景或本体泛化是待测能力，不是获得这一名称的必要条件；本书也不要求每个系统都具有显式世界模型。

1.2 具身不等于“有身体”

带有机械臂并不自动表明模型学会了闭环行为：写死轨迹可能完全不读取反馈，固定脚本也可能通过传感器执行有效控制。应把身体、反馈控制和学习能力分别说明。对机器人系统，本书采用以下检查维度：

- 是否存在闭环反馈，而非一次性输出；
- 行动是否改变未来观测分布；
- 模型是否处理本体状态、几何与动力学约束；
- 是否能从交互、失败或人类纠正中更新；
- 能力是否随场景、任务与本体变化而保持。

1.3 三种尺度：技能、策略与系统

模型论文常只评估中间一层，却用系统级语言描述能力。研究者应主动标注：论文训练和测量的是哪一层，其他层由人、脚本、经典规划器还是隐藏的工程规则完成。

表 1.1 具身研究中经常被混为一谈的三个尺度

尺度	输入与输出	典型时间范围	主要失败模式
低级控制/技能	状态、目标到关节/末端命令	毫秒到数秒	抖动、超调、接触不稳、延迟
通用策略/VLA	图像、语言、本体状态到 action chunk	数秒到数十秒	语义误解、动作多解、分布外状态
完整具身系统	任务到计划、技能执行、验证与恢复	数分钟到数小时	子目标遗漏、记忆丢失、累积误差、安全违规

1.4 形态决定计算问题

轮式导航、单臂、双臂、灵巧手与人形机器人不仅动作维数不同，其接触模式、可达空间、稳定性、传感器与低级控制接口也不同。同一个末端位姿增量，在不同机械臂上需要不同逆运动学解；同一个关节增量，在不同减速器与控制器上产生不同动力学。本体差距（embodiment gap）的讨论强调：共享权重或数据格式不等于目标机器人已经可执行，仍需报告坐标映射、控制器、标定、时序和适配工作 [1]。

1.5 任务先于模型：一个可复述的研究对象

“把杯子放到盘子上”还不是完整实验定义。需要指定机器人与控制器、可见传感器、杯盘初态分布、时间上限、是否允许碰撞和人工接管、终态谓词。例如成功可定义为：目标杯位于指定区域、夹爪释放，且持续一秒不倾倒。这个谓词与“末端到达某点”不同，动作误差小也不能直接推出成功。

同一任务可以分别研究感知、技能或完整系统；若机器人、数据与控制器同时更换，就不能把全部收益归给网络架构。

1.6 一张领域地图

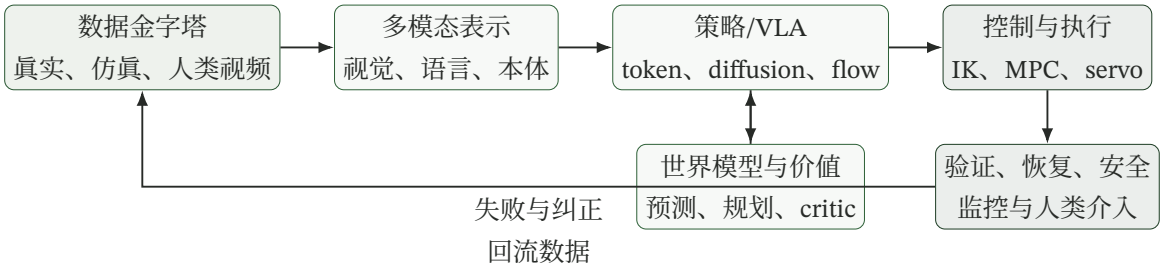


图 1.1 当前具身智能的主要模块与数据闭环。强系统通常是组合，而非单一网络。

本章区分了控制、策略和系统。下一章把这张地图落实为轨迹、信息状态与动作合同。

第二章　数学底座：信息、轨迹与动作接口

本章先固定符号，再把一个二维目标翻译成机器人命令。若轨迹、参考系和时钟没有对齐，后面再精细的训练目标也可能优化错对象。

2.1 状态、观测、历史与信念

机器人任务可写成部分可观测马尔可夫决策过程（POMDP）[2]：

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, \Omega, r, \gamma, \rho_0).$$

状态 s_t 是使动力学满足马尔可夫性质的环境变量，可能含物体姿态、速度、接触、负载与机器人状态。它并不等于传给模型的向量。观测 $o_t \sim \Omega(\cdot \mid s_t)$ 是相机、本体、触觉等可获得的读数；本书把关节位置记为 q_t，把关节、速度、夹爪等拼接的本体向量记为 x_t^{prop}。任务指令 ℓ 在一个回合内视为固定条件，需要时并入历史。

动作 a_t 经控制器改变状态：$s_{t+1} \sim P(\cdot \mid s_t, a_t)$。历史是

$$h_t = (o_0, a_0, o_1, a_1, \dots, a_{t-1}, o_t).$$

策略 $\pi(a_t \mid h_t)$ 能利用历史，反应式策略 $\pi(a_t \mid o_t)$ 只是受限形式。一个网络的隐藏向量不是环境真状态；有限帧历史也不自动具有马尔可夫性质。

信念状态（belief）是给定历史的后验分布 $b_t(s) = p(s_t = s \mid h_t)$，不是一幅图或一个确定标签。在本章观测只依赖当前状态的约定下，贝叶斯过滤分为预测和校正：

$$\bar{b}_{t+1}(s') = \int P(s' \mid s, a_t) b_t(s)\, ds, \tag{2.1}$$

$$b_{t+1}(s') = \frac{\Omega(o_{t+1} \mid s') \bar{b}_{t+1}(s')}{\int \Omega(o_{t+1} \mid u) \bar{b}_{t+1}(u)\, du}. \tag{2.2}$$

离散状态时积分换成求和。精确 belief 在已知模型下是充分信息状态；实际机器人通常用状态估计、时序编码或对象图近似它。近似是否保留决策信息，需要实验检验。

直觉

杯子暂时被手遮住时，当前图像没有杯子位置；历史仍有证据，belief 则表达它可能在哪里及其不确定性。把这三者都叫“状态”，会掩盖感知缺失与策略失败的区别。

2.2 轨迹分布与回报的下标

固定一个包含 T 次动作的回合：

$$\tau = (s_0, o_0, a_0, r_0, \dots, s_{T-1}, o_{T-1}, a_{T-1}, r_{T-1}, s_T, o_T).$$

本章令 $r_t = r(s_t, a_t)$ 为确定的期望奖励；若奖励随机，还应在概率分解中加入其条件分布。轨迹的联合密度为

$$p_\pi(\tau) = \rho_0(s_0)\Omega(o_0 \mid s_0) \times \prod_{t=0}^{T-1} \pi(a_t \mid h_t)P(s_{t+1} \mid s_t, a_t)\Omega(o_{t+1} \mid s_{t+1}). \tag{2.3}$$

这包含终点观测 o_T，并与回报使用相同的动作次数：

$$J(\pi) = \mathbb{E}_{\tau\sim p_\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]. \tag{2.4}$$

若采用无限时域，通常要求 $0 \le \gamma < 1$ 且奖励有界；有限时域也可取 $\gamma = 1$。终止与仅因时间上限截断应分开记录。

行为克隆在专家访问分布上训练，部署则产生自己的轨迹分布。错误不必然逐步爆炸，但在缺乏纠正数据时，策略会访问训练时罕见的历史。第九章说明最坏界所需的假设，以及 DAgger 如何改变数据采样分布 [3]。

2.3 价值与优势属于哪种信息状态

对历史策略，有限时域价值应定义为

$$Q_t^\pi(h, a) = \mathbb{E}_\pi\left[\sum_{k=t}^{T-1} \gamma^{k-t} r_k \mid h_t = h, a_t = a\right], \tag{2.5}$$

$$V_t^\pi(h) = \mathbb{E}_{a\sim\pi(\cdot|h)} Q_t^\pi(h, a), \qquad A_t^\pi(h, a) = Q_t^\pi(h, a) - V_t^\pi(h). \tag{2.6}$$

设 $\bar{r}(h, a) = \mathbb{E}[r(s_t, a) \mid h_t = h]$，则

$$Q_t^\pi(h, a) = \bar{r}(h, a) + \gamma \mathbb{E}_{o_{t+1}|h,a} V_{t+1}^\pi(h, a, o_{t+1}), \quad V_T^\pi = 0. \tag{2.7}$$

精确 belief 可替换历史。只有在状态完全可观测、策略为状态马尔可夫策略等相应条件下，才能直接写 $V^\pi(s)$ 和 $Q^\pi(s, a)$。训练时用仿真真状态的 privileged critic 可以辅助学习，但不能因此声称部署策略看到了真状态。

优势比较的是同一信息条件下的动作价值与当前策略基线，未必是“相对数据常规动作的改进”；优先回放也不等于估计优势。最大熵 RL 的 soft value 还包含熵项，应与这里的普通回报区分。

2.4 动作必须带坐标、单位和控制语义

策略调用率和驱动内环频率不同，例如策略 20 Hz、伺服 1 kHz 是两种时钟。这里的周期是动作被消费的物理周期。增量标签在改变采样率后需要重新组合或换算；绝对目标不能按同样规则相加。

位姿用齐次变换表示：

$${}^{B}T_{E} = \begin{bmatrix} {}^{B}R_{E} & {}^{B}p_{E} \\ 0 & 1 \end{bmatrix} \in SE(3), \quad R \in SO(3),\ p \in \mathbb{R}^3.$$

表 2.1 动作接口示例。频率均为本书教学配置，不是型号规格。

接口	命令及参考系	单位与周期	必须另外声明
关节绝对位置	q^{target}，固定关节顺序	rad；20 Hz	零位、关节限位、插值器与增益
关节增量	Δq 相对本周期目标或测量值	rad/step；20 Hz	相加基准，不能当作 rad/s
末端绝对位姿	${}^{B}T_{E}^{\text{target}}$，基座系 B	m、旋转；20 Hz	工具中心点、IK、可达性
末端相对运动	基座或工具系的平移与旋转增量	m/step、rad/step	左乘或右乘、旋转参数化
速度/力矩	$\dot{q}$ 或关节力矩	rad/s 或 N·m	驱动更新率、饱和与制动规则

相对变换 ΔT 是 $SE(3)$ 的元素；六维小运动向量 $\Delta\xi$ 属于 $\mathbb{R}^6$，通过李代数映射 $\widehat{\Delta\xi} \in \mathfrak{se}(3)$ 和指数映射得到变换。基座系增量一般左乘 $T_{t+1} = \exp(\widehat{\Delta\xi_B})T_t$，工具系增量右乘 $T_{t+1} = T_t \exp(\widehat{\Delta\xi_E})$；同一六个数在两种定义下并非同一运动。若标签只用“平移加法+欧拉角差”，应照实声明，不能暗称为严格的刚体积分。

夹爪命令还要说明是宽度、速度、力还是二值开合，0 表示开还是关，范围是否归一化。旋转须注明角度单位、轴顺序与四元数分量次序；角度环绕和四元数双覆盖会影响普通 MSE。

2.5 二维两连杆到达：同一目标的两种描述

取平面两连杆臂，连杆各长 $l_1 = l_2 = 0.5\,\text{m}$，基座位于原点，$x$ 向右、y 向上，关节逆时针为正，第二角相对第一连杆。正运动学为

$$p(q) = \begin{bmatrix} l_1 \cos q_1 + l_2 \cos(q_1 + q_2) \\ l_1 \sin q_1 + l_2 \sin(q_1 + q_2) \end{bmatrix}.$$

初态 $q_0 = (0, \pi/2)$，故 $p_0 = (0.5, 0.5)\,\text{m}$。目标取

$$q_* = (\pi/6, \pi/3), \quad p_* = (\sqrt{3}/4, 3/4)\,\text{m} \approx (0.433013, 0.750000)\,\text{m}.$$

用**关节增量接口**，总增量为 $(\pi/6, -\pi/6)$ rad；用**末端绝对位置接口**，命令为基座坐标 p_*。IK 的余弦定理给

$$\cos q_{2*} = \frac{\|p_*\|^2 - l_1^2 - l_2^2}{2 l_1 l_2} = \frac{1}{2},$$

选择 $q_2 = \pi/3$ 的肘型，再求 $q_1 = \pi/6$。另一肘型也可能可达，必须由限位、碰撞和种子姿态选解。

总末端增量为 $\Delta p = (-0.066987, 0.25)\,\text{m}$。在初态线性化，

$$J(q_0) = \begin{bmatrix} -0.5 & -0.5 \\ 0.5 & 0 \end{bmatrix}, \quad J(q_0)\Delta q = (0, \pi/12)\,\text{m}.$$

它与精确 Δp 不同，因为 30° 是有限转角，不能把局部 Jacobian 当作精确坐标转换。

为免把几何算例变成不合理的一步跳跃，设 20 Hz 下用 40 步线性关节插值，在 2 s 到达：每步 $\Delta q = (\pi/240, -\pi/240)$ rad，平均关节速度大小 $\pi/12 \approx 0.2618$ rad/s。绝对末端目标由 IK 和同一限速插值器实现时，最终位置相同；若采用笛卡尔直线插值，中间路径则不同。两种动作表示不保证相同碰撞风险。

2.6 预测时域、执行时域与延迟手算

H_o 输入的观测历史长度，同时报告采样间隔，不能只报帧数。

H_p 一次预测的动作步数，预测时域为 $H_p\Delta t$。

H_e 重新获取决策前执行的步数，通常 $1 \le H_e \le H_p$，执行时域为 $H_e\Delta t$。

理想无等待流水线每 $H_e\Delta t$ 重规划一次，频率 $1/(H_e\Delta t)$；同步“执行完再推理”的周期还要加推理与通信等待。预测长度本身不能决定反馈频率。

取 $\Delta t = 50$ ms（20 Hz）、$H_p = 16$、$H_e = 4$。预测覆盖 0.8 s，但每次只承诺执行 0.2 s。曝光/传输 30 ms，推理 120 ms，总可用延迟 $L = 150$ ms。如果在消费旧队列的起点请求新预测，剩余 4 步可覆盖 200 ms，理想余量 50 ms；抖动超过该余量仍会欠载。必要条件是当前有效队列的剩余时间覆盖延迟与抖动，而非简单地用 $L < H_p\Delta t$ 判断。

表 2.2 一次异步预测的教学时间线；动作索引绑定目标执行时刻。

时刻	事件	对齐要求
0 ms	采样，旧队列继续	预测索引 j 对应 $j\Delta t$
30 ms	观测到达	保留曝光时间，不改为接收时间
150 ms	新预测就绪	0、50、100 ms 的动作已过期
200 ms	在边界替换队列	从索引 4 起执行 4 步，不能从索引 0 重播
350 ms	执行该前缀最后一步	观测已老 350 ms；到下一更新可达 400 ms

跳过过期预测需要模型训练时的延迟定义与队列锚点一致；另一选择是显式预测未来起始时刻的动作。表中的连续供给不保证接触安全。若任务要求观测到执行的年龄不超过 100 ms，150 ms 的计算路径已违反要求，缩短 H_e 也无法消除这一下限。

练习与提示

若总延迟改为 420 ms、队列只剩 4 步，缺口为 220 ms。即使 $H_p = 16$ 覆盖 800 ms，仍不能用 4 步旧前缀隐藏延迟。试给出两种可检验改进：降低端到端延迟；提前请求并验证更长缓冲的反馈年龄。增加开环时间能否用于接触阶段，需要单独评测。

2.7 目标与约束是不同的接口

约束任务可写成 $\max_\pi J_r(\pi)$，满足 $J_{c_i}(\pi) \le d_i$。这类期望成本约束并不等于每一步都安全。运行时限速、可达性检查和碰撞过滤可减轻风险，但保证取决于状态估计、模型、延迟、制动与故障假设，第十七章进一步区分规则过滤、控制约束和形式化证明。

第三章　机器人系统栈：端到端模型之外还有什么

3.1　经典模块化栈

一个移动操作机器人常包含：传感器驱动、时间同步与标定、状态估计、语义/几何感知、任务规划、运动规划、轨迹优化、低级控制、碰撞与安全监控。经典栈的优势是接口清楚、可验证、可单独调试；劣势是误差会在硬接口间传播，并且每个模块需要手工建模。

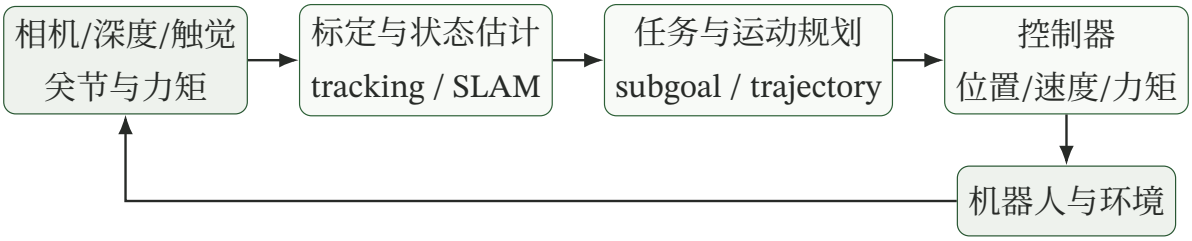


图 3.1　经典机器人闭环。学习模型可以替换某一模块，也可以跨多个模块端到端训练。

3.2　现代学习系统的三种集成方式

模块替换　用神经网络做抓取、视觉伺服、cost map 或 dynamics model，其余保持经典方法。

端到端策略　从图像、语言以及按模型可选的本体状态直接输出低级动作或 action chunk，RT-1、OpenVLA、π_0 属于此类。

层级混合　LLM/VLM 负责高层任务分解，VLA 执行技能，经典规划/控制负责约束，验证器决定继续、恢复或求助。SayCan 用语言先验与可供性价值组合技能得分，是这一方向的经典代表 [4]。

3.3　高层慢思考与低层快控制

高层以技能或子目标为输出，低层以有限时间动作或参考轨迹为输入，驱动内环再以更高频率跟踪。慢与快不是固定的型号数字：自由空间移动、接近接触和插入可能需要不同反馈时限。GR00T N1 等架构使用语义与动作模块分工 [5]，但模型报告的模块吞吐不能直接视为所有传感器到执行器的闭环更新率。

高层提议“抓住杯子”时，低层应收到对象 ID、目标参考系、有效期、速度/力边界和取消句柄。若对象已不可见、夹爪未空或目标过期，执行器应返回可检查的失败码；计划器据此重新观测或换技能。详细技能契约见第八章。

研究提示

“端到端”是训练信号能否跨模块传播，不等于部署时只能有一个网络。一个可训练的 VLA 完全可以外接 IK、阻抗控制、安全 shield 和执行验证；这通常比把所有约束塞进 prompt 更可靠。

3.4 本体状态为何不能省

单幅视觉通常不足以辨识关节速度、夹爪力、执行器饱和与接触。本体向量 x_t^{prop} 可包含关节位置 q_t、速度 $\dot{q}_t$、末端位姿、夹爪宽度、力矩或控制模式。输入哪些量由具体模型与设备决定，不应因统一框图而假定它们全部存在。融合方式包括：

- 线性投影为一个或多个 state tokens；
- FiLM 调制视觉或 action expert；
- 与 noisy actions 一起进入 diffusion/flow transformer；
- 由单独 proprio encoder 产生条件向量。

缺失本体状态有时能在慢速桌面任务中工作，却容易在遮挡、精细插入和动态执行中失败。反过来，训练时过度依赖精确 simulator state 会导致现实不可用。

3.5 标定怎样变成模型可用的信息

相机内参 K 把相机坐标投影到像素，外参 ${}^{B}T_{C}$ 把相机点变到基座坐标。对去畸变像素 (u, v) 和光轴方向深度 $d = Z_C$，

$$p_C = dK^{-1}(u, v, 1)^{\top}, \qquad \begin{bmatrix} p_B \\ 1 \end{bmatrix} = {}^{B}T_{C} \begin{bmatrix} p_C \\ 1 \end{bmatrix}.$$

若传感器给的是射线距离，应先把 $K^{-1}(u, v, 1)^{\top}$ 归一化为单位射线再乘该距离。深度定义、米或毫米的单位都要写入 schema。RGB 与深度还需确认配准；手眼相机的 ${}^{B}T_{C}$ 随关节状态改变，应使用曝光时刻的运动学。若模型只读取 RGB 而不显式使用外参，标定仍通过相机安装、裁剪、视角与训练分布影响策略，并未消失。

在工作空间放置若干已知点，比较反投影与机器人测得坐标的残差，检查尺度、轴向、手眼标定与工具中心点。只在图像平面检查重投影误差不足以证明整个动作链正确：错误的工具偏置仍可能在接触时出现。

3.6 跨模态同步和策略输入

选一个决策锚点 t，从每个传感器取满足时效要求的历史窗口；对本体数据插值时保留原始时间戳与误差范围，离散控制模式和接触事件不宜线性插值。相机曝光、驱动读取、服务器接收、命令下发和实际执行分别计时，跨设备先同步时钟或估计偏差。

策略输入可组织为

$$c_t = F\big(E_v(I_{t-H_o+1:t}), E_p(x_{\le t}^{\text{prop}}), E_f(f_{\le t}), E_l(\ell), m_t, \delta_t\big),$$

其中 m_t 是模态/时间有效 mask，δ_t 是各模态年龄。H_o 帧的实际跨度由采样间隔决定。触觉可在高频窗口内编码后送动作专家；不能把不同频率的数组按行号直接拼接。数据丢失应通过 mask 表达，不应把全零读数冒充“没有接触”。

一个实用诊断是在静态几何、受控运动和接触三种片段中，分别注入固定时间偏移并测闭

环变化。若只改变时间戳就重现失败，先修同步；若无论时序如何都固定向某方向偏移，优先检查外参与工具坐标。

3.7 真正的系统接口清单

复现或构建系统前应固定：

1. 相机数量、分辨率、曝光、内外参与时间戳；
2. observation history 的长度和采样间隔；
3. 动作坐标系、单位、绝对/相对定义与裁剪范围；
4. 控制频率、action chunk 长度、执行多少步后重规划；
5. IK、轨迹平滑、阻抗/位置控制器及其增益；
6. 推理硬件、网络延迟、队列和异步策略；
7. 成功/失败/超时/人工干预定义；
8. 碰撞、力矩、速度、工作空间和急停约束。

本章要点

VLA 不是机器人系统的全部。一个模型的输出需要符合具体驱动合同；末端接口通常还经过坐标变换和 IK，关节或力矩接口则走不同控制路径。队列与约束层仍须可追踪。研究中最常见的“隐藏变量”，正是这些论文正文之外的接口与时序选择。

第二部分

表示与决策：从动作学习到长程执行

第四章　多模态感知与表示

4.1　视觉表示：语义与空间精度的张力

互联网预训练 VLM 擅长类别、语言与常识，但机器人还需要厘米级位置、遮挡、深度、接触前几何和时间变化。常见视觉骨干包括 CNN、ViT、DINO 系列、CLIP/SigLIP，以及带语言解码器的 VLM。OpenVLA 将 SigLIP 的语义特征与 DINOv2 的空间特征拼接，再投影到 Llama 2 的 token 空间 [6]，代表“语义 + 几何细节”的融合思路。

更高图像分辨率会改变 token 数、裁剪与计算预算，却不保证任务成功率提高。OpenVLA 原论文的分辨率消融是特定训练预算与任务下的报告 [6]，不能作为所有精细操作的结论。应先检查相机视角是否包含目标、目标在压缩后是否仍可辨，再比较分辨率与动作噪声等瓶颈。

4.2　二维、三维与多视角

表 4.1　机器人视觉表示的主要选择

表示	优点	代价	更适合
单 RGB 图像	最便宜，可直接复用 VLM	深度歧义、遮挡、历史缺失	语义明显的桌面任务
多视角 RGB	覆盖遮挡，保留预训练兼容性	token/带宽增大，需同步	双臂、移动操作
RGB-D/点云	几何明确，便于 $SE(3)$ 推理	传感噪声、跨设备差异	抓取、插入、空间泛化
体素/三平面	可做全局 3D attention	显存高、量化误差	稠密操作场景
对象/关键点 token	紧凑、可解释、适合规划	检测误差成为瓶颈	结构化长程任务
视频/时序特征	显式运动与状态变化	训练与推理昂贵	动态物体、接触与恢复

PerAct 把 RGB-D 构造成体素表示，用 PerceiverIO/Transformer 预测离散 $SE(3)$ 动作 [7]；Transporter Networks 与 CLIPort 则利用空间相关性，把“拿什么”和“放到哪里”拆开，在二维平面操作中获得强归纳偏置 [8, 9]。这类结构把空间操作的归纳偏置写入模型，是否比通用骨干更合适，需在相同数据与动作合同下检验。

4.3　语言条件究竟提供什么

语言可以是任务 ID 的别名，也可以带来组合、属性、否定、顺序和常识。若训练场景中每条指令都对应唯一背景，模型可能忽略语言而从图像猜任务。要检验真正 grounding，应设计同

一场景多指令、相同对象不同关系、同义改写、未见属性组合和对抗性 distractor。

VIMA 将文字、图像、对象框等统一成多模态 prompt，研究任务组合泛化 [10]；PaLM-E 把连续传感嵌入注入语言模型，展示 embodied multimodal reasoning[11]。但“能回答机器人问题”与“能稳定产生低级动作”仍是不同能力。

4.4 触觉、力与音频

视觉看不到夹爪内部接触和细小滑移。力/力矩、触觉阵列和音频可帮助插入、拧紧、切割和材质判断。RH20T 包含视觉、力、音频、本体等模态与人类示范 [12]；具体使用的模态、数据子集和标定版本应单独报告。融合难点包括不同采样率、接触事件稀疏、传感器跨设备不可比以及触觉标定漂移。

常用做法是把高频信号先用 1D CNN/Transformer 编码，再与视觉 token 交叉注意；或仅在 action expert 中注入，避免 VLM 主干承担所有高频计算。训练时要做时间偏移消融，因为几十毫秒错位即可破坏接触监督。

4.5 从模态编码到动作条件

早期融合把各模态编码为 token 后共同注意，便于跨模态交互，但高频触觉和长视频可能挤占上下文；后期融合先单独提特征，再在动作头结合，较容易控制各通路的速率。交叉注意让动作查询视觉/语言特征，FiLM 则以条件产生缩放和平移。不能只凭框图中的加号认定采用了哪种融合。

语言、视觉与本体信息应各有反事实检查：同一场景换指令测试语言依赖，遮挡目标或换背景测试视觉捷径，在有真实测量的条件下遮蔽本体维度测试其作用。模态 dropout 需要匹配部署缺失方式；训练时全程有触觉、部署时用零填充，不是经过验证的缺失模态适配。

对触觉先去偏、确认力/力矩轴向和单位，再做窗口统计或时序编码；压低接触峰值的均值池化可能同时删掉滑移事件。评测应保留接近接触与稳定抓持两个阶段，分别检查错过事件与误触发，而不是只统计整段重建误差。

4.6 token 压缩与信息瓶颈

RT-1 使用 TokenLearner 压缩视觉特征 [13]。SmolVLA 研究视觉 token 压缩、VLM 层数选择和异步执行 [14]。这些是不同模型版本的工程机制，具体 token 数与速率应随输入分辨率、硬件、批量和版本记录。压缩应回答两件事：删掉的信息是否与动作无关？节省的 FLOPs 是否真的转化为更低端到端延迟？

研究提示

对 VLA 做 token pruning 或量化时，最有价值的指标不是重建误差，而是按任务阶段测量：自由空间移动、接近接触、接触调整、终止判断分别对哪些 token/层最敏感。具身模型的冗余通常是状态依赖的。

第五章　动作生成：回归、token、扩散与流匹配

一个动作头应回答三个问题 :训练时监督什么, 推理时怎样得到物理命令, 计算完成前机器人做什么。本章统一用 c_t 表示由观测历史、指令和可用本体状态编码的条件，$A_t = (a_t, \dots, a_{t+H_p-1})$ 表示归一化后的预测动作序列。归一化统计只从训练集估计，部署时必须反变换。

5.1 回归为何会落在两个有效动作之间

假设绕过障碍可向左或向右，专家在同一观测下等概率给出横向动作 -1 和 $+1$。最小化

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(c,a)\sim\mathcal{D}} \, \|a - f_\theta(c)\|_2^2 \tag{5.1}$$

的最优预测为条件均值 0，而 0 可能正好朝向障碍。这个手算说明单个确定点预测可能不适合未消歧的多峰动作；它不说明所有回归都失败。补充历史或任务条件后，动作可能近似单峰。

高斯策略 $p_\theta(a \mid c) = \mathcal{N}(\mu_\theta(c), \operatorname{diag}\sigma_\theta^2(c))$ 的负对数似然包含加权平方误差和 $\log\sigma$；固定各维方差时才等价于相应权重的 MSE。L1 的最优解是条件中位数，对极端值较不敏感，但也不能自动恢复多个模式。混合密度头、离散 token、扩散和流都可表达更丰富分布，是否值得其计算成本要在同一闭环协议中判断。

OpenVLA-OFT 研究了连续 chunk 与 L1 等微调设计 [15]。这可作为低延迟候选基线，不能从一组任务推导出 L1 普遍优于生成式动作头。

5.2 动作 token：量化、训练与解码

先为每维设定训练集范围 $[l_j, u_j]$，均匀分成 B 个 bin，用中点解码。对范围内、未裁剪的标量，量化误差不超过 $(u_j - l_j)/(2B)$。若平移范围为 $[-0.05, 0.05]$ m/step、$B = 256$，最大中点误差为 0.0001953125 m/step，约 0.195 mm/step；控制器累积、离群裁剪和旋转误差不在这个单步标量界内。

把动作或动作 chunk 编码为 $y_{1:M}$ 后，teacher forcing 目标为

$$\mathcal{L}_{\text{AR}} = -\mathbb{E}_{\mathcal{D}} \sum_{m=1}^{M} \log p_\theta(y_m \mid y_{<m}, c_t). \tag{5.2}$$

训练能并行计算真实前缀下的各位置损失；推理通常逐个产生 token，再反量化为物理量。合法词表 mask 只约束输出语法, 不保证动作可达或无碰撞。RT-2 将末端增量、夹爪与终止等序列化, OpenVLA 基础版按训练动作分位数确定离散范围 [6, 16]；二者的具体动作合同仍须分别读取。

FAST 先将 chunk 变换到频域，再量化并做序列压缩 [17]。它利用时间结构缩短动作编码，但高频接触事件未必同样平滑；码长、重建误差、端到端延迟和闭环结果需要一起报告。自回归模型的多峰表达也受 token 顺序、量化精度、采样和训练覆盖限制。

5.3 Chunk 不是一种独立损失

Chunking 是同时输出未来 H_p 步的时间组织方式，可与回归、CVAE、扩散、流或 token 组合。训练样本应含观测历史长度 H_o、预测长度 H_p、有效动作 mask 和真实采样周期；执行只取 H_e 步再重规划。通常 $H_e < H_p$，所以长预测不必等于长开环执行。

ACT 使用条件变分自编码器（CVAE）预测双臂序列 [18]。训练时后验编码器 $q_\phi(z \mid A, c)$ 看到专家 chunk，目标可写为重建损失加 $\beta D_{\mathrm{KL}}(q_\phi \| \mathcal{N}(0, I))$；推理时不再读取专家未来，而使用规定的潜变量或先验采样。若复现时给模型未来动作，便泄漏了答案。

不同起点可对同一个目标时刻产生重叠预测。可采用一种本书定义的时间融合：

$$\bar{a}_t = \frac{\sum_{k=0}^{K} e^{-\lambda k} \hat{a}_t^{(t-k)}}{\sum_{k=0}^{K} e^{-\lambda k}}, \qquad \lambda \geq 0, \tag{5.3}$$

其中 k 越大表示计划越老，因此本式偏向新预测，不声称复刻 ACT 原代码的加权索引。只有同一目标时刻、同一参考系与可线性混合的量才能直接相加；旋转和二值夹爪需相应融合规则。接管、接触突变或子目标切换时应让旧计划失效。

第二章已区分吞吐、反馈年龄和队列。第十八章通过实际运行的教学仿真演示：增加预测兼执行长度可能仅稍微改变离线误差，却显著改变受扰动后的闭环成功率。这是所选环境的联合时域效应，不是所有 chunk 模型的普遍规律。

5.4 Diffusion Policy：训练一个去噪器

Diffusion Policy 把条件扩散用于动作序列 [19]。以下给出 DDPM 形式以说明机制 [20]，具体实现可使用不同采样器。令干净动作是 $A^{(0)}$，噪声等级 $k = 1, \ldots, K$，$\beta_k \in (0, 1)$，$\alpha_k = 1 - \beta_k$，$\bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i$。训练可直接采样

$$A^{(k)} = \sqrt{\bar{\alpha}_k} A^{(0)} + \sqrt{1 - \bar{\alpha}_k}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

并最小化

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{A^{(0)}, c, k, \epsilon} \left\| \epsilon - \epsilon_\theta(A^{(k)}, k, c) \right\|_2^2. \tag{5.4}$$

直觉是：每次训练随机遮蔽不同程度的动作结构，模型依靠条件与剩余结构识别噪声。它没有要求图像也扩散；视觉骨干可每次决策只计算一次。

选择使 $\bar{\alpha}_K$ 足够接近 0 的噪声日程，终端扰动分布才能近似标准高斯；否则直接使用高斯先验会失配。推理从 $A^{(K)} \sim \mathcal{N}(0, I)$ 开始，一个常见反向步为

$$A^{(k-1)} = \frac{1}{\sqrt{\alpha_k}} \left(A^{(k)} - \frac{\beta_k}{\sqrt{1 - \bar{\alpha}_k}} \epsilon_\theta(A^{(k)}, k, c) \right) + \sigma_k z,$$

其中 $z \sim \mathcal{N}(0, I)$，最后一步常取 $z = 0$；σ_k 必须与采用的采样器约定一致，不能任意替换。反向迭代结束后反归一化、检查动作合同、只执行有效前缀，再读取新观测。少步采样可降低延迟，却改变生成近似误差；需要重新测量任务表现。

Listing 5.1 条件扩散的训练与执行接口伪代码

```
# Training: episode split and normalization stats are frozen.
cond, clean, valid = sample_training_window()
k = sample_noise_level()
eps = normal_like(clean)
noisy = sqrt(abar[k]) * clean + sqrt(1-abar[k]) * eps
loss = masked_mean((eps_model(noisy, k, cond)-eps)**2, valid)
update(loss)
# Inference: K reverse steps, then use only a time-aligned prefix.
cond = encode_available_history()
chunk = normal_actions(H_p)
for k in reversed(range(1, K+1)):
    chunk = reverse_step(chunk, k, cond)
execute(denormalize(chunk), steps=H_e, check_timestamp=True)
```

扩散能表达多峰不等于每次能选对任务模式。它可能在低覆盖条件下生成不可执行序列，迭代延迟也可能让条件在执行时过期。归一化、动作边界、末端速度与夹爪事件仍由接口负责。

5.5 Flow matching：学习从噪声走向动作的速度

条件流匹配（flow matching）可用线性路径介绍 [21]。用 $u \in [0,1]$ 表示**生成过程的时间**，它不是控制时刻 t。对独立噪声 ϵ 和专家动作 A，令

$$A^u = (1-u)\epsilon + uA, \qquad \frac{dA^u}{du} = A - \epsilon.$$

训练目标为

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{A,c,\epsilon,u} \left\| v_\theta(A^u, c, u) - (A - \epsilon) \right\|_2^2 . \tag{5.5}$$

平方损失的最优速度场是给定 A^u, c, u 的目标速度条件均值，而非记住每一对独立噪声和动作。推理时从 $A^0 \sim \mathcal{N}(0, I)$ 积分 $dA^u/du = v_\theta(A^u, c, u)$，例如 N 步 Euler 更新

$$A^{u+\delta} = A^u + \delta v_\theta(A^u, c, u), \qquad \delta = 1/N.$$

即使训练路径是直线，学习到的边缘速度场的轨迹也不必逐样本直线；有限步积分、网络误差和数据覆盖都会影响结果。减少 N 需同时测延迟与闭环质量。π_0 使用 VLM 条件和独立动作专家实现流式连续 chunk 生成 [22]，不同论文可能反转噪声/数据的时间端点，复现时必须连同速度符号一起换算。

5.6 按失效条件选择动作头

比较时固定历史、动作表示、数据划分、控制器和执行时域。把单步回归与执行很长前缀的扩散直接比较，测到的是多个因素的合成。若要研究动作头本身，应把 H_o, H_p, H_e 及硬件时序一并控制。

表 5.1 动作头的条件与诊断。多峰表达能力不是闭环优越性的排名。

路线	可取的条件	典型失败	最小诊断
单点回归	条件较充分，动作近单峰，延迟紧	模式平均、尺度失衡	同观测动作分布；各维物理误差
自回归 token	复用序列主干，码长可控	裁剪、量化、逐 token 等待	码长/频率/误差按任务阶段统计
扩散	需要多峰连续序列，迭代预算够	条件过期、采样不稳	固定数据与种子扫采样步数
流匹配	连续动作与独立专家接口匹配	积分误差、条件错配	统一端点方向并扫求解步数

练习与提示

回归例子的均值为 0，为什么混合高斯能保留左右两种选择？提示：区分分布建模与从分布采样。再设流匹配目标为 $A^u = u\epsilon + (1-u)A$，写出目标速度与推理起点；只改公式中的 u 而不改积分方向，会得到什么？

本章要点

回归学习点或显式密度，token 学习离散序列条件分布，扩散学习去噪过程，流匹配学习连续速度场。Chunk 可用于所有这些路线；最终执行的前缀、观测年龄与重规划频率须独立说明。

第六章　VLA 与机器人基础模型的架构谱系

6.1 先比较接口，再比较主干

前一章把动作输出拆成回归、离散 token 与连续生成。本章进一步问：模型实际看见什么，学习的是哪种机器人的动作，论文提供什么证据？这些问题比总参数量更接近复现的起点。一种允许缺失模态的抽象是

$$u_t = F_\theta\big(E_v(I_{t-H_o+1:t}), E_l(\ell), E_p(x_t^{\text{prop}}), m_t\big), \qquad A_t \sim D_\theta(\cdot \mid u_t).$$

这里 u_t 是网络表示，不能与历史 h_t 或 belief b_t 混用；x_t^{prop} 是本体状态，m_t 是模态有效性、时间或本体标识等元数据。并非每个模型都输入历史和本体状态。$A_t = (a_{t|t}, \dots, a_{t+H_p-1|t})$ 预测 H_p 步，实际执行 H_e 步后才重新取得策略输出，二者必须另外登记。

以下按随书保存的论文版本核对，资料核查截至 2026-10-22。本书没有重训这些大型模型，也没有重做论文眞机试验；参数、数据规模和消融均为论文自报。“公开证据”说明论文及项目入口能支持什么，不能视为本次已复现。复现时还需冻结代码、权重、数据许可和接口版本。

6.2 从机器人专用策略到网页语义共训

6.2.1 RT-1：压缩视觉 token 后预测动作

观测 最近 6 张 RGB 图像与语言。FiLM 调制 EfficientNet，TokenLearner 压缩视觉 token，再交给 Transformer。

动作 末端平移/旋转、夹爪、底盘及模式分别离散；连续量使用 256 个 bin。模式选择移动底盘、操作手臂或终止，不能将全部输出视为同一个 7 维接口。

数据与本体 约 130k 次遥操作示范，来自 13 台同型移动操作机器人、17 个月及 700 余条任务指令。同型实例变化不等于任意跨本体迁移。

证据与版本 arXiv:2212.06817v2（2023-08-11）[13]，报告约 35M 参数、3 Hz 系统及约 3,000 次眞机试验；完整数据和 fleet 系统不能由论文单独重建。

其数据多样性消融只支持所测配方，不能推出扩大任务宽度在任何情况下都胜过增加示范次数。

6.2.2 RT-2：将动作写进 VLM 输出词表

观测 图像与语言任务，主干来自 PaLI-X 或 PaLM-E；基础策略不能默认拥有任意长视觉记忆。

动作 将末端 6 自由度位移、夹爪量与终止命令序列化为文本 token；连续量各离散到 256 个 bin，解码限制为合法动作 token。

数据与本体 网页图文与机器人轨迹共微调，主要控制 RT-1 同类移动操作平台。网页知识不替代目标动作示范与控制器适配。

证据与版本 arXiv:2307.15818v1（2023-07-28）[16]，比较至 55B 的模型，在指定对象、符号和指令协议中报告迁移；完整权重、数据和训练设施未随论文开放。

合法 token 约束输出语法，没有保证末端可达、动作不碰撞或任务完成。

6.2.3 Open X-Embodiment 与 RT-X：数据和模型分开记录

观测 OXE 保存各源图像、语言和可用本体信息；RT-1-X、RT-2-X 采用各自支持的通道，不能无条件使用 schema 中全部字段。

动作 RT-X 重点对齐末端平移、旋转与夹爪的 7 维表示；归一化、原坐标和控制频率仍需逐源核查。

数据与本体 发布快照汇集 21 个机构、22 类机器人、百万余条轨迹。OXE 是数据集合，RT-1-X、RT-2-X 是在相应混合数据上训练的策略。

证据与版本 arXiv:2310.08864v9（2025-05-14）[23]，提供数据/转换入口和跨本体消融；上游数据许可与各模型权重开放范围需分别核查。

统一容器方便 loader 读数据，物理语义仍须对齐：相同 Δx 在工具系和基座系中一般指向不同方向。

6.3 开放策略与连续动作专家

6.3.1 Octo：保留主干，适配传感器与动作头

观测 图像历史与语言或目标图像，block-wise attention 和模态掩码支持输入变化。新增腕相机或本体状态是适配选项，不是全部预训练配置的既有输入。

动作 readout token 读取观测与任务，条件 diffusion head 预测 chunk；预训练混合筛选末端增量控制数据，迁移至关节位置等空间需对应动作头与数据。

数据与本体 25 个 OXE 数据集、约 800k 条轨迹，多种机械臂/相机/场景；提供 27M 与 93M 预训练模型。

证据与版本 arXiv:2405.12213v2（2024-05-26）[24]，公开代码与 checkpoint，报告零样本/微调结果。适配样本量、时间与硬件是任务条件，不能作为统一保证。

新增状态通道也可能引入虚假相关，应与通道缺失、时间错位、视觉遮挡对照，不能仅因输入更多推断性能更好。

6.3.2 OpenVLA：单步离散动作与开放微调

观测 基础版本输入单张 RGB 图像与语言，不输入本体状态；DINOv2 与 SigLIP 特征接 Prismatic/Llama 2 主干。

动作 6 维末端增量加 1 维夹爪，共 7 维，逐维离散后自回归输出。每维按训练数据第 1–99 百分位区间划分 256 个 bin；超范围裁剪，反归一化须选择正确数据源统计。终止量不能从 RT-2 接口直接移植到本模型。

数据与本体 约 970k 条 OXE 策展轨迹、多种机械臂；共同末端接口没有消除控制器、频率和坐标差异。

证据与版本 arXiv:2406.09246v3（2024-09-05）[6]，公开 7B 模型与微调代码；视觉微调、分辨率、LoRA 和量化结论来自指定消融。

例如论文的 BridgeData V2 部署采用 5 Hz 非阻塞控制器，在 RTX A5000 上比较精度：8 位实现约 1.2 Hz，4 位约 3 Hz。成功率同时受数值误差和时序影响，不能推出“4 位普遍无损”或“8 位普遍更差”；本次未测这些 GPU 配置。OpenVLA-OFT 后续改为并行连续 chunk 等配方 [15]，应与原 token 模型分开；原始 OpenVLA 没有独立连续 action expert。

6.3.3 π_0：VLM 连接 flow 动作专家

观测 多相机图像、语言与本体状态；VLM 初始化自 PaliGemma，状态和含噪动作还进入动作专家。

动作 约 300M 专家用条件 flow matching 生成连续 chunk，论文取 $H_p = 50$；关节/末端适配随机器人变化，50 步不等于固定秒数。

数据与本体 约 10,000 小时自有数据，覆盖 7 种配置、68 个任务，并混合 OXE；特定任务另用高质量数据后训练。

证据与版本 arXiv:2410.24164v4（2026-01-08）[22]，论文及 OpenPI 提供机制和模型入口；公开子集不能完全替代自有训练分布。

chunk 命令间隔的高频不代表每步重新运行 VLM；缓存、H_e 和推理期间续接规则须另报。

6.3.4 $\pi_{0.5}$：异质共训与分层指令

观测 图像、语言、本体状态；先预测高层语义子任务，再条件化低层动作，高低层频率不同。

动作 训练结合 FAST 离散动作任务与连续 flow matching，部署由连续专家输出 chunk；训练 token 任务并非全部是运行时必经步骤。

数据与本体 移动操作、静态机器人、网页视觉语言和高层子任务等异质来源，用于家庭移动操作泛化。

证据与版本 arXiv:2504.16054v1（2025-04-22）[25]，报告家庭任务/共训消融，OpenPI 提供入口；完整自有数据和家庭协议仍限制独立复现。

6.3.5 $\pi_{0.7}$：将行为方式写入条件

观测 多视角图像、MEM 风格视频历史、本体状态和子任务，可附加质量/策略元数据及子目标图像。

动作 约 860M flow 专家生成连续动作，模式区分关节/末端命令；条件指定某种方式不等于该方式必然实现。

数据与本体 多本体轨迹、次优/失败片段、人类视频与非机器人任务，按速度、质量和错误等元数据标注。

证据与版本 arXiv:2604.15483v2（2026-04-24）[26]，图 2 记录 4B Gemma3 VLM、历史编码与专家，总计约 5B；公开材料不能重建完整自有分布。

该机制适合研究混合质量数据的条件建模，不支持直接无条件模仿失败轨迹。

6.4 动作专家、双系统与小模型的其他选择

6.4.1 RDT-1B：为异构物理量保留位置

观测 多视角图像历史、语言、本体状态与控制频率，后者区分动作对应的时间尺度。

动作 1.2B diffusion Transformer 生成连续 chunk；将关节、末端和夹爪等量填入统一空间，用掩码表示有效分量。

数据与本体 46 个数据集、百万余条预训练轨迹，再以目标双臂机器人 6k 余条示范微调；统一向量仍需本体适配。

证据与版本 arXiv:2410.07864v2（2025-03-01）[27]，提供论文、模型和代码入口；双臂结果不能代表其他本体的零样本能力。

6.4.2 CogACT：语义表示与动作生成组件化

观测 当前图像与语言，VLM 形成 cognition feature；基础实现不能默认拥有任意状态记忆。

动作 DiT 条件化该特征生成连续末端动作序列，以 adaptive ensemble 融合重叠 chunk；最大动作模块约 300M。

数据与本体 OXE 预训练与目标域适配，评测包含 Google Robot、Realman 与 Franka 等设定。

证据与版本 arXiv:2411.19650v1（2024-11-29）[28]，约 7B VLM 加动作模块，提供模型/代码入口；增益须连同 SIMPLER 或真机协议理解。

扩大动作模块可能改善表达，也可能增加迭代延迟，应在固定数据和实际时序下消融。

6.4.3 DexVLA：动作预训练、本体对齐与新任务适配

观测 多视角图像与语言，Qwen2-VL 产生动作条件与推理 token。各实验状态通道以配置核对，不能因机器人有编码器就认定已输入策略。

动作 最大约 1B diffusion expert，多本体输出头映射各自动作空间。连续专家与推理 token 是不同学习对象。

数据与本体 约 100 小时跨本体示范，采用专家预训练、本体对齐、新任务后训练三阶段，涉及单臂、双臂与手等设置。

证据与版本 arXiv:2502.05855v3（2025-08-09）[29]，论文与项目展示课程和多本体实验；阶段收益依赖初始化与数据，不能概括为课程学习必然更好。

6.4.4 GR00T N1：本体相关编码器连接共享主干

观测 图像、语言及本体状态；Eagle-2 VLM 与本体相关状态/动作编码器连接。

动作 DiT 用 action flow matching 预测 chunk，本体相关解码器处理各自动作维度；不同机器人的关节编号不能直接混合。

数据与本体 真实遥操作、OXE、人类视频、仿真和生成轨迹。88 小时遥操作至 827 小时生成视频是特定合成管线规模，生成视频不等于独立真实交互。

证据与版本 arXiv:2503.14734v2（2025-03-27）[5]，公开 GR00T-N1-2B 约 2.2B 模型和工具；合成及真机效果均受论文协议限定。

6.4.5 Gemini Robotics：云端语义与本地动作解码

观测 场景图像、任务语言与本体信息；偏空间推理的 Robotics-ER 与直接控制的 Robotics 应分开记录。

动作 云端 backbone 加本地解码器输出连续 chunk。远端响应、端到端延迟和 chunk 命令频率是不同量。

数据与本体 ALOHA 2 遥操作及网页/具身推理数据，另适配双臂 Franka 与 Apollo；手部等接口随本体变化。

证据与版本 arXiv:2503.20020v1（2025-03-25）[30]，报告推理、适配和真机能力；论文未提供重建全套策略的公开权重与完整训练数据。

6.4.6 SmolVLA：异步系统也是模型设计的一部分

观测 多相机图像、语言与传感运动状态，经 SmolVLM2 和状态投影编码。

动作 连续 flow expert 生成 chunk；异步队列让执行与预测并行，但须丢弃过期预测并对齐起点。

数据与本体 从 481 个社区数据集选择约 22.9k episodes、10.6M 帧，并在仿真/低成本真机适配；较小预训练集合不代表目标数据无关。

证据与版本 arXiv:2506.01844v1（2025-06-02）[14]，约 450M 参数，提供代码、模型和数据入口；速度及成功率须连同硬件和队列实现理解。

6.5 从模型卡选出基线

表 6.1 先冻结问题，再比较设计轴；不构成成功率排名。

问题	首先固定	可比较的变量
动作头是否合适	观测、数据、动作单位	回归、token、生成头；H_p, H_e
预训练是否迁移	目标数据与预算	初始化、冻结范围、适配模块
跨本体是否成立	留出本体与适配成本	共享/专属编码器、归一化
部署是否有效	硬件、控制器、截止时间	精度、缓存、实际延迟

练习与提示

选择一个模型，写出实际输入、动作反归一化和执行频率，再画一次观测到命令生效的时间线。提示：找不到 H_e、相机时间戳或控制器模式时，先列为复现缺口，不能从一个“Hz”猜出全部时序。

本章要点

模型谱系提供设计选择及其证据范围。观测、本体、动作、数据、控制时序和公开资产共同决定能否建立公平基线；参数量或跨协议成功率不能替代这些记录。

第七章　世界模型、规划与想象

7.1　机器人规划需要什么后果预测

世界模型预测动作介入后的变化。若真实状态可用，可学习 $\hat{s}_{t+1} = f_\theta(s_t, a_t)$；部分可观测时，规划器通常从历史 h_t 提取内部表示 $z_t = \phi(h_t)$，再预测 $p_\theta(z_{t+1} \mid z_t, a_t)$。这个学到的 z_t 不自动等于 belief $b_t = p(s_t \mid h_t)$，也未必是充分统计；是否丢失接触、对象身份或执行器滞后，需要用任务检验。潜变量后验、RSSM 与表征学习的完整推导见《世界模型》中关于信念状态、潜变量动态和模型学习的讨论。本章只保留规划器与机器人接口之间必要的内容。

World Models、MuZero、DreamerV3 和 TD-MPC2 提供不同接口实例：压缩观测后学习动态、为奖励和价值保留隐状态、在想象轨迹中训练策略、在潜空间中滚动优化 [31–34]。它们的训练目标、动作空间和环境不同，不能仅因都使用 latent 就互换算法。对机器人而言，模型的交付物应包括输入时间戳、状态更新、动作语义、预测范围、终止/约束输出以及不确定性说明。

7.2　MPC：规划多步，只执行有限步

给定 z_t，模型预测控制（MPC）求一段候选动作：

$$
\begin{aligned}
A_t^* = \quad & \underset{a_{t|t},\dots,a_{t+H_p-1|t}}{\arg\max} \left[\sum_{k=0}^{H_p-1} \gamma^k \hat{r}(z_{t+k|t}, a_{t+k|t}) + \gamma^{H_p} \hat{V}(z_{t+H_p|t}) \right], \\
& z_{t|t} = z_t, \quad z_{t+k+1|t} = f_\theta(z_{t+k|t}, a_{t+k|t}), \\
& a_{t+k|t} \in \mathcal{A}, \quad g(z_{t+k|t}, a_{t+k|t}) \le 0.
\end{aligned}
\tag{7.1}
$$

这里 H_p 是规划步数，$\mathcal{A}$ 限制动作，g 表示模型能够检查的约束。若在预测中到达终止态，后续奖励及终端价值按任务定义置零；终端价值应与折扣、奖励尺度和规划域一致。未知碰撞或模型未表示的物理量不会因为公式中写了 g 就获得保证。

CEM（交叉熵方法）从候选分布采样动作序列，滚动模型，保留高分样本，再更新均值/方差，多轮迭代后选出候选。它是有限预算下的近似随机优化，不保证全局最优；方差过早坍缩还会失去探索。可用 VLA 的 action chunk 初始化候选，减少无意义搜索，但仍应保留受约束的扰动与失败处理。执行前 $H_e \le H_p$ 步后重新观测，且规划、网络和命令生效延迟必须计入动作起点。若计算结果已经过期，不能把 $a_{t|t}$ 当作当前命令直接追加。

直觉

设两个候选 chunk 的预测距离都减少 5 cm，但一个会让末端穿过桌面。先检查可行性，再用预测任务收益排序，比先最大化视觉相似度再补救更明确。即使排序正确，接触发生后也需用新观测重规划；旧的开环预测不能代替接触反馈。

7.3　视频预测、动作 grounding 与逆动力学

Genie 从无动作标注视频学习可交互环境 [35]，其中 latent action 是模型内部控制变量，不是某台机器人的关节/末端命令。将生成视频用于机器人，还要建立观测投影、时间尺度和本体

动作之间的映射。同一像素运动可以由相机移动、物体移动或不同关节运动造成；视觉未来通常不能唯一确定控制命令。

VERA 的随书版本 arXiv:2605.27817v1（2026-05-27）保持视频规划器不变，为不同本体训练 inverse dynamics model，并利用本体 Jacobian 连接视频运动与动作 [36]。这支持“共享视频规划器、分别适配动作接口”的研究思路。跨视角、不可见接触、不可达视频子目标和数据中没有的动作仍是接口缺口。

预测空间也会影响模型保留什么。Mask World Model 的版本 arXiv:2604.19683v2（2026-04-22）预测语义 mask 动态并连接 diffusion 动作头 [37]。它在论文的 LIBERO、RLBench 和真机设置中报告结果；不能据此认定语义 mask 对所有接触任务充分。如果目标依赖细微触觉、摩擦或材料形变，去掉外观也可能丢失关键线索。

7.4 把组合系统变成可检查的循环

世界模型可以给 VLA 候选排序、生成子目标、提供未来表征或辅助构造训练数据。这些用途应分开消融：未来表征带来的收益不等于规划带来的收益，合成轨迹也不能当作独立真机证据。一个有限执行的接口可写成：

```
history = append_timestamped_observation(history, observe())
state = update_state(history)
candidates = propose_vla_chunks(state, goal)
feasible = check_mode_units_limits(candidates)
plan = rank_with_model(feasible, horizon=H_p)
if plan is None or plan.is_stale():
    fallback_and_log()
else:
    execute_bounded(plan, steps=H_e)
    verify_and_log(observe(), plan)
```

验证器应检查对象/接触/任务状态，不只比较 RGB 重建误差。例如用预留正常轨迹标定关键点误差的阈值，在固定压力测试上报告误报和漏报；采用连续多次超阈触发与冷却规则，避免反复切换。若该信号未校准，应把它当作诊断，而非可靠风险概率。

研究提示

世界模型评价至少有三层：预测误差、候选动作排序是否正确、真实闭环任务是否改善。图像质量指标更好不保证后两层；测试还要防止优化器找到模型漏洞。固定真实环境预算，比较无模型 VLA、模型排序与错误模型排序，才能判断收益是否来自后果预测。

本章要点

机器人世界模型需要回答“在当前接口下执行这段动作会怎样”。规划时域、实际执行步数、状态更新、约束和失败处理必须连成一个循环，预测能力才有机会变成闭环收益。

第八章　长程具身 Agent：计划、技能、记忆与恢复

8.1 把长任务拆成可验证的有限执行

长程任务包含对象跟踪、阶段切换、终止判断和失败恢复。设 S_i 为第 i 段成功，要求不经恢复连续完成时，

$$P(S_1 \cap \cdots \cap S_n) = \prod_{i=1}^{n} P(S_i \mid S_1, \ldots, S_{i-1}).$$

若这些条件成功率都等于 p，才得到 p^n。取 $p = 0.95, n = 20$ 得约 0.3585，这是理想化手算，不是实际长程系统的预测。现实中错误相关、状态越来越难、重试和恢复都会改变该值；单技能平均成功率不足以推断整段任务成功率。

8.2 SayCan：语言相关性与当前可执行性

SayCan 在有限候选技能集中组合语言得分与技能价值 [4]：

$$\text{score}(k \mid h_t, g) \propto p_{\text{LM}}(k \mid g, \text{text}(h_t))\, \hat{p}_k(\text{success} \mid b_t). \tag{8.1}$$

语言项只接收目标和已执行技能等文字历史 $\text{text}(h_t)$，衡量下一技能是否相容；后一项利用机器人状态估计当前可执行性。此乘积是选择启发式，不是联合概率定理；模型分数的尺度、校准与候选集覆盖会影响排序。若价值函数不是成功概率，应明确映射，不能直接当作可供性概率。若全部技能不可执行，候选集应允许停止、重新观测或求助，不能强迫选择分数最大的危险动作。

8.3 一个可审计的技能契约

自然语言名称只能表达意图。以“把杯子放入托盘”为例，技能还需要对象身份、坐标、有效时限和验证规则。

表 8.1　技能接口应返回可检查字段，具体阈值由平台标定。

字段	示例与作用
参数与版本	cup/target 对象 ID、schema/策略版本，避免同名对象混淆
动作契约	末端增量、基座系、米/弧度、控制周期、H_p, H_e
前置条件	目标可见、夹爪持有杯子、目标区空闲；未知也作为一种结果
执行边界	工作空间、速度/力限制、资源锁、取消与超时
后置条件	杯子位于托盘且夹爪已释放，绑定图像/状态证据及时间戳
结果与恢复	success/timeout/occluded/lost-object 等失败码，重观测/退回/求助

检查前置条件后，执行器只取得有限时域的命令权限；过期、取消或断连触发明确 fallback。结果可为成功、失败或未知。遮挡导致验证未知时，不应自动当作成功或重新执行可能重复的动作。运行时修改或 shield 投影后的实际命令也必须记录，不能只保存原策略输出。

EmbodiedSkills 的版本 arXiv:2609.01281v1（2026-09-01）把技能决策作为执行提议，检查前置条件、限制低层执行并验证结果 [38]。论文分别报告低层任务适配和记忆依赖长程任务，不能用前者成功率替代后者；本次没有复现该框架。

8.4 记忆要能更新，也要允许不知道

视觉短期记忆用于速度、遮挡和接触变化；状态记忆保存对象与任务谓词；情节记忆供检索历史失败；程序记忆保存技能与恢复规则。四者不必全部放进 VLM 上下文。一个状态条目至少记录

$$(\text{对象}, \text{谓词}, \text{取值/未知}, \text{证据时间}, \text{来源}, \text{有效条件}).$$

“抽屉已打开”若仅来自五分钟前的图像，在有人介入后应失效。新证据与旧记录冲突时需重观测或保留多假设，而不是检索到相似文本就覆盖当前状态。这种结构化记忆是 belief 的工程近似，置信分也未必校准。

8.5 验证、恢复与评测必须共用同一条轨迹

闭环可按“更新历史与 belief—选择子目标—检查契约—有限执行—验证—恢复或重规划”运行。恢复也占时间和动作预算，人工接管之后的成功要单独标记。预先冻结最大尝试次数和超时，不能每失败一次就重置初态或从统计分母删除。数据可区分原始成功、恢复成功、人工辅助成功、最终失败，并保存首次失败原因与后续事件。

失败轨迹可用来学习检测、恢复或价值判断；不应把全部失败动作当作专家动作。RoboMIND 提供失败与原因标注的实例 [39]，但标注覆盖范围和质量仍需逐源检查。分层评测应同时报告任务最终成功、无干预成功、恢复代价和验证误判。

边界与误区

“我已经完成”不是任务证据。验证应绑定对象区域、关节状态、夹爪力或任务谓词等观测；谓词的传感器误差、遮挡和时效也应接受测试。

练习与提示

设计“打开抽屉后取出物体”的三种轨迹：开屉成功、拉动但未打开、打开后被人关上。写出技能结果、记忆更新与下一步动作。提示：第三种轨迹检验状态失效规则；第二种检验命令执行与目标实现的区别。

本章要点

长程 Agent 需要可审计的状态与技能契约。有限执行、独立验证和不重置分母的恢复记录，连接了强低层策略与可信的整段任务证据。

第三部分

训练与部署：从示范到交互

第九章　行为克隆、分布偏移与交互式模仿

前面的动作头回答“怎样表示并生成动作”，本章回答“应在什么数据上学习”。同一个网络，既可拟合专家示范，也可拟合策略出错后的纠正片段；这两种数据覆盖的状态往往比网络层数的差异更重要。

9.1　行为克隆：在历史上做条件预测

设任务语言为 ℓ，截至当前的可用历史为 $h_t = (o_{0:t}, a_{0:t-1})$；本体向量 x_t^{prop} 是观测的一部分，不能与完整物理状态 s_t 混同。实际模型通常只保留最近 H_o 步，并非总能由这段有限历史恢复 Markov 性。专家数据应保存完整 episode，训练时再抽取窗口：

$$\mathcal{D}_E = \{(h_t, \ell, A_t, m_t)\}, \qquad A_t = (a_t, \dots, a_{t+H_p-1}).$$

其中 $m_{t,j} \in \{0,1\}$ 表示第 j 个未来动作是否真实存在。$H_p = 1$ 时退化为单步行为克隆（behavior cloning，BC）。最大似然目标为

$$\mathcal{L}_{\text{BC}} = -\mathbb{E}_{(h,\ell,A)\sim\mathcal{D}_E} \log p_\theta(A \mid h, \ell). \tag{9.1}$$

离散动作或动作 token 用交叉熵；连续高斯头在固定方差下对应平方误差，Laplace 头在固定尺度下对应绝对误差。损失形式也是对条件分布的假设。多种有效动作的平均未必有效，因此低均方误差并不自动意味着可执行的行为。

以逐位置回归头为例，带掩码的损失可写成

$$\mathcal{L}_{\text{mask}} = \frac{\sum_{t,j} m_{t,j} \|\hat{a}_{t+j} - a_{t+j}\|_1}{\sum_{t,j} m_{t,j}}, \qquad \sum_{t,j} m_{t,j} > 0.$$

这里每个动作向量的各维已按已声明的物理尺度归一化。若还存在缺失动作维度，应另加维度掩码；不能把“不存在的关节”填零后当作真实零动作监督。扩散、flow 或自回归 chunk 头也需要相同的有效性信息，但其训练目标应按模型分别构造。

BC 无需在线探索，容易利用已有示范，是许多机器人策略的起点。它能否恢复错误，取决于数据是否覆盖错误后的状态及正确纠正动作；“成功 episode”内部可以包含恢复片段，“失败 episode”内部也可以包含值得学习的行为。

9.2　误差为何复合：上界及其假设

先考虑一个便于推导的有限时域例子：每步代价 $c_t \in [0,1]$，未折扣累计代价为 $J_c(\pi) = \mathbb{E}_\pi \sum_{t=0}^{T-1} c_t$。专家和学习者从同一初态出发，专家分布上的单步动作不一致概率至多为 ϵ。用相同环境随机性耦合两条轨迹；第一次动作不一致以前，它们访问相同状态。由并集界，在前 $t+1$ 步

至少出现一次不一致的概率至多为 $(t+1)\epsilon$。若偏离后最坏情况下每步都多付出一个单位代价，则

$$J_c(\pi) - J_c(\pi_E) \le \sum_{t=0}^{T-1} \min\{1, (t+1)\epsilon\} \tag{9.2}$$

$$\le \min\left\{T, \frac{T(T+1)}{2}\epsilon\right\}. \tag{9.3}$$

这解释了常说的 $O(T^2\epsilon)$ 最坏情况量级 [3]。它不是每个控制系统都会达到的下界，更不是把连续动作 MSE 代入 ϵ 后就成立的定理。反馈控制可能纠正小偏差；相反，接触中的小位移也可能触发不可逆失败。由动作距离推出任务代价，需要额外的动力学敏感度和恢复能力假设。

训练损失估计的是专家访问分布上的错误；部署访问分布由学习者自己的动作决定。仅增加同一路径上的重复示范，未必补上“抓偏后该怎么办”的覆盖缺口。

9.3 DAgger：让专家标注学习者会访问的历史

DAgger（数据集聚合）在训练中改变采样分布 [3]。下面给出历史条件策略的接口版本：

1. 由初始专家数据训练 π_1，设置聚合数据集 $\mathcal{D}$。
2. 第 i 轮以混合策略 $\pi_i^{\text{mix}} = \beta_i \pi_E + (1-\beta_i)\pi_i$ 收集 episode；此处的混合指每个决策步随机选择行为策略，并记录实际执行动作。β_i 随训练衰减。
3. 请专家为采到的历史给出纠正标签，将 (h_t, ℓ, a_t^E) 加入 $\mathcal{D}$；专家标签与实际执行动作分开保存。
4. 在聚合数据上更新学习者；用独立验证 rollout 选择策略，测试集不参与选择。

理论上的改进依赖可查询且一致的专家、监督学习的低遗憾性质以及动作错误对应的后续额外代价。在专家额外代价可由常数 u 控制等条件下，可以得到形如 $O(uT\bar{\epsilon})$ 的策略代价界；若 u 随 T 增长，便不能笼统声称总是线性改善。

真实机器人还要满足采样安全与专家可标注性。人类接管、纠正演示、仿真 DAgger 和 rollout 后离线标注都是相关采集方式，但它们的选择偏差、标注延迟和专家覆盖不同，不能直接继承完整 DAgger 保证。若专家依赖学习者看不见的真状态，相同观测历史可能对应冲突标签；数据聚合也不能消除这种信息缺失。

9.4 GAIL：匹配哪些占用分布

对于完全可观测、折扣因子 $0 \le \gamma < 1$ 的 MDP，定义归一化折扣状态动作占用

$$\rho_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_\pi(s_t = s, a_t = a).$$

连续空间中应把它理解为测度或密度。GAIL 训练判别器区分专家与策略样本，并用判别反馈优化策略 [40]。本书约定 $D(s,a)$ 为“来自专家”的概率：

$$\min_\pi \max_D \; \mathbb{E}_{\rho_E}[\log D(s,a)] + \mathbb{E}_{\rho_\pi}[\log(1 - D(s,a))] - \lambda \mathcal{H}_\rho(\pi), \tag{9.4}$$

其中 $\mathcal{H}_\rho(\pi) = \mathbb{E}_{\rho_\pi}[-\log \pi(a \mid s)]$；熵正则权重吸收了归一化常数。于是策略最小化 $\log(1-D)$，等价于追求奖励 $-\log(1-D)$。采用“D = 策略概率”的资料会交换两项，这只是约定不同。

对固定 π，理想判别器为 $D^* = \rho_E/(\rho_E + \rho_\pi)$；代回无熵项的目标，得到常数加两倍 Jensen–Shannon 散度。这说明理想优化在匹配占用分布，实际有限数据和有限容量并不保证达到该解。占用相同也不意味着逐条轨迹或隐藏状态转移完全相同。

部分可观测任务可以让判别器输入历史表示、观测动作或 belief；此时匹配的是所选表示的占用，不能声称已匹配不可见的真实状态。GAIL 通常需要环境交互和强化学习更新，样本成本、判别器过拟合及奖励投机都需要诊断；它与直接最大化专家动作似然解决的问题不同。

9.5 序列 BC、数据增强与重标注

Chunk 监督能学习专家的局部时间结构，但只执行前 H_e 步、再用新观测规划时，闭环表现同时取决于 H_p, H_e, H_o 和推理延迟。训练窗口不能跨 episode；末尾 padding 不能参与动作损失；历史 padding 也要有掩码，避免模型把补零当真实观测。

不同频率的数据应先还原物理时间与动作含义。位置目标可按时间插值；相对位移应按时间区间组合，旋转需按其几何表示组合，夹爪开合和接触事件通常是离散事件。简单复制标签或仅统一数组长度会改变行为。

视觉裁剪、颜色扰动和背景替换应保持任务对象可见，几何增强要同步变换相机与动作。语言改写必须保持对象指代、左右关系和目标一致。历史丢帧、延迟扰动可用于增强时序鲁棒性，但应在验证中检查是否损失精细控制。

目标重标注（goal relabeling）将未来实际到达的状态作为条件，用已有轨迹构造新的到达问题。Octo 使用语言或目标图像条件及相应的条件缺失训练 [24]；这类方法增加可学习的条件关系，不会自动产生到达原先未覆盖状态的示范。对于失败或人工接管数据，还需标记行为来源，避免把危险动作当正例。

9.6 从离线诊断走向闭环实验

至少同时观察按物理单位还原的单步误差、多步 rollout、任务分层成功率、接管次数、恢复率与动作时序。训练误差下降而成功率下降时，先检查数据划分、动作逆变换、历史因果性和控制频率，再判断是否需要更强模型。

第 十八 章用轻量任务固定 episode 划分、训练预算与随机种子，比较观测历史或执行 chunk 的影响。它用于验证分布偏移与反馈机制，不能代替真实机器人证据，也不用于给 VLA 架构排名。

练习与提示

令 $T = 20, \epsilon = 0.01$，计算式 (9.3) 的上界；再令 $T = 200$，解释为何必须保留 $\min\{T, \cdot\}$。若只有归一化动作 MSE 为 0.01，还缺什么条件才能作同样推断？

提示：先区分错误概率与动作距离，再考虑一次错误能造成多少后续额外代价。

第十章　从离线数据到强化学习后训练

示范告诉策略“数据中的行为是什么”，奖励和成本定义“哪些后果值得追求”。当已有策略能完成部分任务时，部署失败、执行时间和恢复过程可以成为改进信号。本章只介绍理解机器人后训练所需的强化学习接口；一般算法理论可结合《强化学习》继续学习。

10.1　先定义决策单位与信息条件

策略可以每个控制周期输出动作，也可以每 H_e 步生成一个 chunk，或由高层选择持续若干秒的技能。这些是不同的决策过程。若一个决策跨过 d 个低层周期，应累计这段奖励，并用 γ^d 对下一决策的价值折扣；不能把每个 chunk 当一“步”却保留未解释的逐周期奖励和折扣。

下面 SAC 与 IQL 的公式先在完全可观测 MDP 中用 s 表示状态。部分可观测任务应使用充分 belief、历史条件价值或明确的循环近似；一张图像或任意编码向量并不天然成为 Markov 状态。若直接把不足的观测当状态，Bellman 目标的偏差也是需要测量的模型误差。

10.2　最大熵目标与 SAC

SAC 在任务回报之外加入策略熵 [41]。对于固定动作坐标与概率密度约定，

$$J_{\mathrm{soft}}(\pi)=\mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1}\gamma^{t}\big(r_t-\alpha_{\mathrm{ent}}\log\pi(a_t\mid s_t)\big)\right].$$

$\alpha_{\mathrm{ent}}>0$ 控制回报与熵的权衡；连续动作为微分熵，其数值依赖动作单位和尺度，因此归一化与温度约定必须一起记录。熵奖励鼓励分布更分散，并不保证探索安全。

实际常用两个 critic Q_{ϕ_1},Q_{ϕ_2} 及其目标网络。令 d_t^{term} 表示任务真正终止，采样 $a'\sim\pi_\theta(\cdot\mid s')$，目标为

$$y=r+\gamma(1-d^{\mathrm{term}})\left[\min_{j=1,2}Q_{\bar{\phi}_j}(s',a')-\alpha_{\mathrm{ent}}\log\pi_\theta(a'\mid s')\right]. \tag{10.1}$$

这里采用将 soft 值函数 $V_{\mathrm{soft}}(s')=\mathbb{E}_{a'\sim\pi}[Q(s',a')-\alpha_{\mathrm{ent}}\log\pi(a'\mid s')]$ 展开后的目标；早期 SAC 实现另有显式学习的 V 网络，不应把两种实现混作逐行相同的算法。两个 critic 分别拟合 $(Q_{\phi_j}(s,a)-y)^2$，更新 critic 时对 y 停止梯度；目标网络缓慢跟踪在线参数。策略用可微重参数化等方式最小化

$$\mathcal{L}_\pi=\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\theta}\left[\alpha_{\mathrm{ent}}\log\pi_\theta(a\mid s)-\min_j Q_{\phi_j}(s,a)\right].$$

这里 $\mathcal{D}$ 是回放数据，温度也可另行学习。时间上限造成的截断与真正终止要分开：若截断只是采样停止而任务仍可继续，通常需要 bootstrap；若有限任务时域本身就是目标定义的一部分，应把剩余时间纳入状态并在终点终止。错误合并两者会改变价值目标。

QT-Opt 展示了视觉抓取中大规模离策略 Q-learning 的可行性 [42]，但其系统还依赖自动复位、机器人试验和数据基础设施。算法名称相同，不代表实验室能以同样成本复现完整闭环。

10.3 离线 RL：策略为何会利用数据外高估

离线数据固定时，Bellman 更新和策略优化可能查询数据没有支持的动作。若 critic 在这些动作上高估，策略就会追逐估计错误，进而加剧高估。约束策略贴近数据、惩罚数据外价值以及避免显式的数据外最大化，是不同的应对思路；它们都需要足够的任务相关覆盖。

IQL 用数据动作学习 expectile 值函数，再做优势加权模仿 [43]。定义

$$L_2^\tau(u) = |\tau - \mathbf{1}[u < 0]|u^2, \qquad \frac{1}{2} < \tau < 1,$$

以及目标 critic $\bar{Q}$。一个简化的训练目标组为

$$\mathcal{L}_V = \mathbb{E}_{(s,a)\sim\mathcal{D}} L_2^\tau\big(\mathrm{sg}[\bar{Q}(s,a)] - V_\psi(s)\big), \tag{10.2}$$

$$\mathcal{L}_Q = \mathbb{E}_{\mathcal{D}}\Big[\big(Q_\phi(s,a) - \mathrm{sg}\big[r + \gamma(1 - d^{\mathrm{term}})V_\psi(s')\big]\big)^2\Big], \tag{10.3}$$

$$w(s,a) = \min\{w_{\max}, \exp\big(\beta\,\mathrm{sg}[\bar{Q}(s,a) - V_\psi(s)]\big)\}, \tag{10.4}$$

$$\mathcal{L}_{\mathrm{AWBC}} = -\mathbb{E}_{\mathcal{D}}\big[w(s,a)\log\pi_\theta(a \mid s)\big]. \tag{10.5}$$

sg 表示停止梯度，$\beta > 0$ 是逆温度，$w_{\max}$ 防止少量大权重主导更新。双 Q、目标网络和数值稳定化属于实现配置，也应记录。τ 越大，V 越偏向数据内较好的动作；这仍受数据与函数逼近限制。

IQL 的价值学习避免对显式数据外候选动作作最大化，但训练后的策略仍可能输出数据外动作。优势加权也不能修复错误价值、缺失的恢复行为或完全没有覆盖的任务。把它用于 VLA，需要先确认奖励、任务尺度、历史表示和动作头能支持相应更新。

10.4 机器人后训练与语言模型后训练的关系

“预训练、监督微调、强化学习后训练”描述的是训练阶段，不指定完整控制算法。用语言模型术语讨论机器人时，至少要补齐表 10.1 的接口。

表 10.1 把后训练配方落到机器人前必须回答的问题

项目	语言序列中的常见对象	机器人中需另行定义的对象
决策与概率	token 及其可计算条件概率	控制动作、chunk 或技能；连续密度或隐式生成路径
奖励与信用	回答整体或过程评分	成功谓词、接触、安全成本、持续时间和低层信用分配
数据来源	已采样回答与偏好	示范、自主 rollout、接管、控制器修改和复位过程
策略更新	概率比、KL 或分组相对奖励	行为策略、动作尺度、离策略校正及可优化的动作似然
执行约束	输出文本的约束	推理延迟、频率、队列、可达性、碰撞和硬件限制

例如，PPO 概率比需要在同一动作定义下计算新旧策略概率；动作 token 的交叉熵可以计算，不代表一个 flow 生成的最终连续 chunk 的精确密度同样易得。组内相对奖励可以成为控制

任务的估计器，但还需要明确分组任务、采样策略、优势和方差处理，不能把 GRPO 名称直接等同于连续控制算法。

一个可审计的流程是：先用 BC 在目标机器人建立可工作的初始策略；固定任务与成功判据；收集自主失败和接管记录；验证奖励及价值模型；选择能与动作头兼容的保守更新；最后在独立回归任务上检查成功率、安全和时延。数据足够时可比较优势加权 BC、离线 RL 和受限在线更新，避免一次同时改变数据、动作头和奖励。

截至 2026-10-22，已有工作报告 fleet 部署学习、任务分组后训练与细粒度在线适配 [44–46]。本书用它们说明机制选择，不把各自任务上的论文自报成绩合成为统一排名；闭源数据、机器人资源与奖励设计会限制结论的迁移。

10.5 奖励、成本与接管日志

稀疏成功奖励可来自环境谓词、传感器或人工判断；几何进度奖励需要可信状态；VLM 或人类偏好评分适合补充语义，但应检查视角、遮挡、提示和评分偏差。过程奖励能缩短信用链，也可能让策略优化评分器而非最终目标。

安全成本应单独记录。把碰撞赋一个有限负奖励，不等于保证永不碰撞；可采用约束目标

$$\max_{\pi} J_r(\pi) \quad \text{s.t.} \quad J_c(\pi) = \mathbb{E}_{\pi} \sum_{t} \gamma^t c_t \le C,$$

但估计约束满足也不等于逐时刻、逐次运行的硬保证。训练目标、运行时约束和验证证据需要分别说明。

人类接管既改变数据分布，也改变实际执行策略。日志应分别保存策略建议、shield 后命令、最终执行动作和行为来源。否则奖励可能把人类纠正的后果错误归给原策略，或让 critic 学到没有实际执行的转移。接管成功不能记作原策略自主成功。

10.6 扩散与 flow 策略怎样接受后果反馈

多步去噪或积分得到的 chunk 通常具有隐式分布。可选路线包括：对示范或候选 chunk 做优势加权回归；把生成步骤视为内部决策过程；用 Q 引导采样；或对速度场施加可微的价值相关更新。每种方法都要说明优化的是最终动作分布还是内部生成过程，近似引入了什么偏差。

只更新动作专家或 adapter 可以降低计算开销，但不保证稳定，也可能无法修正视觉误解。对初始策略的行为约束可减少偏离，约束强度仍须按任务验证。训练成功率上升时，还应检查动作幅度、长程退化、奖励投机和延迟。

练习与提示

一个技能持续 $d = 4$ 个周期，每周期折扣为 0.99。写出该技能的一次 Bellman 目标，说明为什么下一状态价值前的系数应为 0.99^4。若技能中间被人类接管，哪些动作和奖励还能归属于原行为策略？

提示：先确定决策时刻，再写累计区间奖励；接管后的数据可以保留，但行为来源必须明确。

第十一章　数据混合、适配与跨本体学习

广覆盖数据提供状态和任务的多样性，目标域数据帮助策略适应具体相机、身体与行为要求。二者有时冲突：不同操作者可能采用不同路径，动作接口可能不一致，质量标签也可能缺失。预训练、共训练和后训练的划分，应服务于这些实际差异。

11.1　预训练与后训练分别解决什么

预训练通常希望学习可迁移表示及较宽的行为支持；后训练针对目标任务、行为质量、部署资源或奖励改进。π_0 等工作采用覆盖广的预训练与更有针对性的后训练数据 [22]，这是一种数据组织选择，不能推出所有任务都应采用同样阶段和比例。

恢复能力来自对偏离状态及有效纠正的覆盖。只包含顺利完成路径的数据，可能缺少恢复；包含多种策略的数据若有条件标签或能表示多峰的动作头，也可以学到一致行为。数据“完美”或“混杂”不是足够的诊断，应该检查状态覆盖、行为来源、标签和目标之间的关系。

11.2　采样权重与损失权重不要混为一谈

设第 i 个数据源为 $\mathcal{D}_i$，采样概率为 w_i，源内已按有效动作或 token 归一化的损失为 $\bar{\mathcal{L}}_i$。再用 λ_i 调整不同任务目标的相对权重：

$$\mathcal{L}(\theta) = \sum_i w_i \lambda_i \, \mathbb{E}_{z \sim \mathcal{D}_i}[\bar{\mathcal{L}}_i(z;\theta)], \qquad w_i \geq 0, \quad \sum_i w_i = 1. \tag{11.1}$$

改变 w_i 也改变一个源被重复访问的次数，改变 λ_i 只改变其梯度贡献；有限训练预算下，两者不总等效。若一个源按帧计样本、另一个按长 episode 计样本，直接比较 n_i 也会误导。

一种温度采样启发式是 $w_i \propto n_i^{\eta}$，$0 < \eta < 1$，其中 n_i 必须采用一致采样单位。它降低大源的支配程度，但并不保证最优。也可按任务、场景、本体分层采样，或上采样高质量目标数据。依据梯度大小动态配权要检查是否放大噪声，课程训练要检查后加入的数据是否导致遗忘。

实现要点

比较数据配方时固定总优化步数、batch 的有效动作数和目标域验证规则。记录每源被采样次数、重复率、损失尺度、梯度范数与闭环指标。新增数据的收益可能来自更多训练步数或更多目标样本，必须用对照分开。

Octo 和 OpenVLA 的数据配方表明，多源数据需要按内容和可学习程度筛选 [6, 24]。某个训练预算下移除数据有收益，不代表该数据集普遍有害；不同模型容量、动作对齐和训练时间可能改变结果。

11.3　共训练：哪类信息能传到控制

机器人轨迹具有直接动作标签；通用图文带来语言与语义覆盖；人类视频提供行为和对象交互线索；仿真数据提供可控状态、动作与失败覆盖；子任务语言和目标图像帮助条件化。它们

并非等价样本，缺失的本体动作或时序不能靠统一 token 格式自动补齐。

截至本书核验日期，共训练研究已比较图文、跨本体轨迹、人类视频和动作表示的组合 [47]。阅读结论时应问：总机器人数据是否固定，是否增加了训练预算，目标能力是否由对应测试集测量，语言能力与控制能力是否分别评估？一项模态在某个任务上的收益，不支持“模态越多越好”的普遍判断。

小规模实验可以先固定机器人数据和训练预算，只改变一个辅助源；测量它应改善的泛化轴，同时检查原任务退化。若结果无提升，应检查对齐、标签噪声和损失比例，再决定是否扩大规模。

11.4 冻结、LoRA 与全量微调的选择

表 11.1 适配策略及其需要验证的条件

策略	直接收益	主要限制	应检验的条件
仅训练动作头	更新参数少	不能补回冻结表示丢失的细节	特征是否保留位姿、接触与目标语义
冻结部分主干	降低训练开销	域差异可能发生在冻结模块	新视角、材质和语言分别在哪里失效
LoRA / adapter	低显存、便于保存多任务增量	插入层、秩和训练数据影响结果	与相同数据下的头部及全量基线比较
全量微调	允许全部模块共同适配	显存、遗忘与过拟合成本较高	数据是否支撑自由度，旧任务是否退化
分阶段适配	可分开检查动作与语义对齐	更多阶段与超参数	各阶段是否提供独立可验证收益

OpenVLA 的视觉冻结与 LoRA 实验、DexVLA 的动作专家课程提供了具体比较 [6, 29]；其中的秩、显存、训练时间和接近全量微调的结论，都应放回各自数据与硬件条件。全量微调拥有更多可更新参数，不构成效果上限必然更高的保证。

选择模块前先定位差距：新相机导致空间误差、语言模板导致目标误解、本体变化导致动作不可达、还是执行队列导致迟到？只有最后一层动作不匹配时，仅训练动作头才有充分动机。

11.5 跨本体对齐：共享数组之前先共享含义

常见路线包括统一末端位姿及夹爪、带有效性掩码的关节槽位、本体专属动作头、对象中心动作、共享潜动作加本体解码器，以及高层计划配本体控制器。RDT、DexVLA 和视频规划加本体逆动力学分别展示了这些方向的部分实现 [27, 29, 36]。

一个动作字段的对齐契约至少包含：

1. 本体型号、关节顺序、左右臂顺序、工具中心点及标定版本；

2. 坐标系、长度和角度单位、旋转表示、绝对或相对定义；
3. 控制模式、周期 Δt、最大速度/位移、底层插补与夹爪语义；
4. 有效维度、缺失传感器、归一化统计与目标本体适配数据；
5. 可达性、碰撞、关节限位、控制器失败与回退规则。

共享 7 维“末端动作”不表示身体相同。不同机械臂的冗余、工作空间、负载、夹爪和低层控制器都可能改变动作后果。

11.5.1 相对位姿的左右乘就是不同接口

设 ${}^{B}T_{E,t}$ 把工具坐标中的点映射到基座坐标。工具局部增量与基座增量分别可定义为

$$ {}^{B}T_{E,t+1} = {}^{B}T_{E,t}\Delta T_{E}, \qquad {}^{B}T_{E,t+1} = \Delta T_{B}\,{}^{B}T_{E,t}. $$

同一数值矩阵一般产生不同目标；若要表达同一个刚体运动，二者满足 $\Delta T_{B} = {}^{B}T_{E,t}\Delta T_{E}({}^{B}T_{E,t})^{-1}$。还应说明采用完整 $SE(3)$ 组合，还是将平移与旋转分别按某一坐标系相加；二者不是自动相同的实现。

11.5.2 运动学映射与物理可执行性

在已声明坐标系和刚性运动学模型中，关节速度与末端瞬时 twist 的关系为

$$ \xi = J(q)\dot{q}. $$

这是该运动学模型中的精确瞬时微分关系；将有限关节位移直接映射为有限末端位移时，才使用局部线性近似。逆映射需要处理冗余、奇异点、关节限位和碰撞；即使几何目标可达，速度、加速度、力与接触也可能不可行。把本体动作映射到统一空间后，还要通过真实控制器验证反向映射。

所谓跨本体零样本应分别报告：策略权重是否更新、动作头是否训练、是否使用目标数据统计、是否重新标定、是否重写控制 wrapper。只共享主干不等于系统没有目标域适配。

11.6 训练稳定性与验证顺序

先在小样本上过拟合，核查坐标、掩码、逆归一化与行为标签；再扩大数据源。可按视觉、语言主干与动作头设不同学习率，并记录梯度裁剪、混合精度和优化器配置。哪些层需要更高数值精度，应由数值与闭环对照支持，不能只按模块名字断言。

验证集必须隔离用于主张泛化的场景、任务或本体；统计量只从训练集估计。训练期间同时监控每源损失、闭环成功、旧任务回归与端到端延迟，避免“总体损失降低”掩盖小数据源的退化。

练习与提示

两套三维末端增量中，A 用工具系毫米，B 用基座系米且频率翻倍。仅统一单位再归一化是否足够？列出还需的处理。提示：坐标变换依赖姿态；每步位移的物理意义也随周期改变。

第十二章　Sim2Real：辨识差距，再选择适配方法

仿真适合可重复初态、自动奖励、并行试验和失败诊断。它的价值来自可控实验，而真实转移取决于仿真遗漏了什么：相机、标定、接触、摩擦、柔性、执行时序与控制器都可能形成差距。本章把 Sim2Real 视为一组可区分的系统假设。

12.1　域随机化的目标与覆盖条件

设环境参数为 ϕ，包括光照、纹理、相机、质量、摩擦、执行延迟和控制增益。域随机化从训练分布 $p_{\text{train}}(\phi)$ 采样并优化

$$\max_{\pi} \ \mathbb{E}_{\phi \sim p_{\text{train}}(\phi)}[J(\pi;\phi)].$$

经典工作展示了视觉随机化促进转移的可能 [48]。其直觉是让策略不要依赖某个纹理或精确参数，但平均性能目标并不保证每个参数组合都安全或成功。即使真实参数处于训练范围内，也可能因优化不足、稀有组合、未建模现象和观测差异而失败。

随机化应对准因果因素。仅改变颜色不足以覆盖相机外参偏差、通信抖动或接触刚度；过宽的随机化也可能让任务不可辨识，迫使策略学习过分保守的行为。报告参数范围、采样分布、相关关系与随机种子，而不仅写“使用了 domain randomization”。

12.2　系统辨识先缩小不确定性

用独立的真实校准日志估计控制器响应、时延、噪声和关键动力学范围，再在可信区间训练。可执行的顺序为：

1. 先核查单位、坐标和相机内外参，排除纯接口错误；
2. 在允许的安全范围内执行可辨识的命令，比较命令与位置、速度或力响应；
3. 对延迟、增益、摩擦等拟合候选模型，保留不确定区间而非只给单点；
4. 在仿真中分别与联合扰动这些参数，用保留的真实验证日志检查预测；
5. 固定配方，再在未用于辨识与调参的真实测试场景评估。

互相关可用于有足够激励、时间基准一致的信号延迟诊断；在强闭环、缺乏激励或动作与观测非线性关联时，它可能给出错误解释，需要结合时戳和系统辨识。

12.3　残差学习：保留基控制器的条件

在已定义的动作坐标和单位下，可让学习器修正基控制器：

$$a_t = \text{Proj}_{\mathcal{A}_t}\big(a_t^{\text{base}} + \Delta a_\theta(h_t, \ell)\big).$$

投影集合 $\mathcal{A}_t$ 应包含当时有效的动作限制。只有两项代表同一接口时，加法才有物理意义；对位姿还应按相应群运算或已声明的小量参数化组合。残差可聚焦于模型误差或接触调整，但它不必降低动作维数，也不会自动继承基控制器的稳定性。

例如，位置增量残差若持续把目标推到限位，低层控制器虽然每步都正确跟踪，系统仍可能失败。需要限制残差幅值、变化率与适用区域，验证饱和及退出条件；安全投影本身的局限见第 十七 章。

12.4 数字孪生与合成示范

Real-to-Sim-to-Real 把真实几何、对象状态和任务逻辑带入仿真，生成新初态或诊断试验，再回到真实验证。RoboMIND 的数字孪生与 RoboCasa 的厨房任务提供了相关资源 [39, 49]。视觉相似只是其中一项：可交互关节、碰撞网格、摩擦、夹爪接触与成功谓词仍需检查。资源规模依赖论文与发布版本，本书不把不同版本中的场景数、对象数和轨迹数拼成一个数据量。

MimicGen 将源示范按对象相关片段进行变换并重新执行 [50]，其有效性依赖对象对应、可达性、规划与成功筛选。视频或神经生成数据可以补充视觉与行为线索 [5]，但视频时长、预测动作时长和真实机器人执行时长是不同单位。

对合成数据逐项检查：

- 动作标签是否来自可执行控制，还是从图像估计的代理标签；
- 碰撞、接触与对象状态是否物理一致，成功过滤是否可靠；
- 源示范和生成副本是否全部留在同一个数据 split；
- 数据是否只扩增已有轨迹的外观，是否增加任务所需的状态覆盖；
- 混入后是否在固定真实测试集和相同训练预算下改善闭环表现。

合成失败数据也有价值，但学习目标必须区分“覆盖失败状态”和“模仿失败动作”。

12.5 把转移失败分解成最小诊断

表 12.1 Sim2Real 差距与可区分的诊断；先验证接口，再判断模型不足

差距	典型症状	最小区分试验
視觉	对背景、光照或遮挡敏感	固定眞实物理状态，改变成像因素；有配对眞值时测感知误差
几何与标定	呈固定方向或尺度偏移	已知位姿点阵测量残差，检查外参和工具中心点
动力学与控制	超调、滞后、振荡	同一受限命令比较状态响应，分别改变负载和增益
接触	插入、滑动或抓取失效	隔离接触子任务，比较力曲线、滑移与接触时刻
时序	离线预测好，执行迟到	记录曝光到执行时间，注入可控延迟并检查队列
状态覆盖	首次偏离后持续失败	在两域施加同类受控扰动，比较恢复率和恢复数据

“眞实图像加仿眞眞状态”只有在获得可靠配对或可控重建时才是有效试验；不能把两域不对应的记录直接拼接。一次改变多个因素可能改善结果，却不能定位原因。诊断集允许探索原因，正式测试集在固定配方后使用；看到测试失败再调仿眞参数，应重新建立未参与调参的测试集。

练习与提示

策略在仿眞中抓取成功，眞机总向相机右侧偏 2 cm，但换光照影响很小。设计两个最小试验区分标定误差和动力学滞后，再说明哪一种情况下扩大纹理随机化没有直接动机。
提示：比较不同速度的静态目标跟踪，并检查用标定板测得的几何残差。

本章要点

转移依赖视觉、几何、动力学、接触、时序与覆盖共同成立。随机化提供分布训练，辨识缩小参数不确定性，残差与眞实微调补具体缺口；它们的收益都应在隔离调参的眞实测试协议中验证。

第十三章　训练与部署中的时序、归一化和执行

一个预测正确却迟到的动作，也可能使机器人失败。本章把训练样本和部署队列放在同一时间轴上：先核查模型看见什么、学了什么、何时执行，再讨论计算优化。

13.1　从曝光到执行的时间戳

每个传感器记录曝光或采样时间 t_{exp}，而非只记录回调到达时间。另记数据接收、推理开始/结束、命令发送和控制器实际应用时间 t_{apply}。不同设备需要共同时间基准或已估计的时钟偏差；本机持续时间宜采用单调时钟。

一个动作的观测年龄为

$$\Delta_{\text{age}} = t_{\text{apply}} - t_{\text{exp}}.$$

它包含成像、传输、预处理、排队、推理和命令执行等待。仅测 GPU 前向时间会漏掉其中大部分环节。多相机和本体传感器还需记录各自时间戳与匹配误差，不能把“同一个 batch”当作同时采样。

训练中的 (h_t, a_t) 应按明确的决策与执行约定配对。若标签来自未来时刻，必须说明是在预测未来控制还是在对齐延迟；错误错配可能让模型学习特定设备的补偿，在换硬件后失效。历史窗口只能使用决策时已可取得的信息，不能因离线重采样引入未来帧。

13.2　动作归一化：统计、逆变换与饱和

对连续动作维度，可用训练集统计定义

$$\tilde{a}_j = \text{clip}\left(\frac{a_j - m_j}{s_j}, -c, c\right), \qquad s_j > 0.$$

m_j, s_j, c、单位和适用数据源必须随模型保存。统计只能由训练 split 估计，不能提前查看测试分布。不同本体未经物理对齐时，不宜共享同一统计。

以分位数映射为例，令 $m_j = (p_{99,j} + p_{1,j})/2$、$s_j = (p_{99,j} - p_{1,j})/2$，取 $c = 1$。OpenVLA 的分位数离散范围提供了相关实例 [6]，具体分位数、离散边界与夹爪处理仍需按实现记录。若某平移维的训练范围为 $[-0.02, 0.03]$ m，则 $m = 0.005$ m、$s = 0.025$ m，预测 $\tilde{a} = 0.4$ 还原为 0.015 m。若模型被 clip 到 1，它只能还原到上边界，不能恢复原先超出的动作。

零尺度或近零尺度应采用已声明的下限或常量维规则，不能除零。缺失维度用掩码；二值夹爪和终止另设分类或已定义的连续接口；角度环绕与四元数约束应在几何表示中处理，不能仅靠逐维标准差修复。记录训练标签和预测的饱和比例，区分模型失误与裁剪后失真。

13.3　预测时域、执行时域和异步队列

H_o 是观测历史长度，H_p 是每次预测的动作步数，$H_e \le H_p$ 是新反馈前实际执行的步数，控制周期为 Δt。模型生成长序列，不代表必须开环执行整段；重规划频率也不等于低层控制频率。

若请求新计划时队列尚有 q 个可用动作，一个基本的不断流条件是

$$L_{\text{request}} + L_{\text{margin}} < q\Delta t,$$

其中请求延迟包括前后处理、通信和推理。还需满足长期吞吐，并检查延迟抖动；这不是反馈稳定性保证。队列足够长可以避免空等，却可能继续执行基于旧观测的动作。

异步服务器与机器人客户端的解耦可减少等待 [14]。每个动作应带计划 ID、依据的观测时间、目标执行时刻和失效条件。新 chunk 到达时：

1. 对齐到当前执行时间，丢弃已经过去的前缀，不能从其第一个动作重新播放；
2. 根据接口决定替换未执行后缀或融合；相对位移与旋转不能盲目求平均；
3. 检查新计划与当前状态的一致性，必要时重规划；
4. 接管、接触突变或安全触发后立即使旧计划失效；
5. 队列欠载或断连时进入已测试的本地保持、受限回退或停止，记录事件。

低层“保持”也可能在接触中施力，停止规则必须与控制模式和硬件配合。

13.4 自适应执行长度是一项待验证策略

执行长度可依据风险评分 u_t、视觉变化、接触事件或进度下降调整。例如

$$H_e(t) = \operatorname{clip}\left(\lfloor H_e^{\max} - k u_t \rfloor, H_e^{\min}, H_e^{\max}\right), \qquad 1 \le H_e^{\min} \le H_e^{\max} \le H_p,$$

其中边界为整数，k 的量纲与评分定义一致。这是启发式，u_t 未校准时不能解释为失败概率。预测步数 H_p 可以保持不变，改变的是开环执行时间 $H_e\Delta t$。

自由运动可能容忍长执行段，精细接触可能需要更快反馈；缩短执行段也会增加推理请求，造成新的排队。验证应在相同数据、模型与硬件下比较成功率、观测年龄、欠载次数和推理负载，而不仅比较平均动作误差。

13.5 缓存、编译和压缩如何影响闭环

若多次生成迭代共享图像与语言条件，可缓存特征或固定前缀，前提是该缓存确实不依赖当前噪声动作、时间条件或待更新状态。π_0 的结构提供了相关实现例子 [22]。编译、算子融合、视觉批处理和异步预处理的收益取决于形状、设备、内存与实际调度；训练分片工具的作用不能直接当作在线推理加速。

量化和 token 压缩应比较最终端到端时延及成功率。视觉细节、时间嵌入、小幅动作和输出层可能有不同敏感性，需按任务阶段测量。通用量化理论与部署实现见《智能体模型量化压缩与部署》；本书重点是压缩怎样改变接口误差、反馈频率与闭环后果。

13.6 终止、恢复与人工接管

终止既可以由策略预测，也可以由独立成功验证器或环境规则判断。过早终止导致未完成，过晚动作可能破坏成功状态。需要声明成功是否保持若干周期、是否允许重试，以及超时如何处理。进度或不确定性头只是辅助证据，不能凭模型自报替代外部判据。

人工接管记录至少包含接管前状态、触发原因、操作者、起止时间、策略建议与纠正动作、恢复结果。接管期间应停止旧队列，重新交还控制权前验证状态与计划。第 十六 章区分自主、辅助和恢复指标。

13.7 一个可审查的训练循环接口

先核查小样本拟合、动作逆变换、坐标、单位、掩码与历史因果性，再比较训练和部署的相机、控制器及频率。接口通过后，才改变覆盖、混合权重、动作头、时域或容量；每次改变一类原因并保留固定初态的失败记录。

下面是接口伪代码，模型、数据加载器和损失需另行实现；可运行实验见第 十八 章。

Listing 13.1 区分动作头目标与验证模式的训练接口

```
model.train()
for step in range(num_steps):
    batch = sample_training_batch(split_manifest)
    cond = encode_history(batch, train_stats)
    valid = batch.action_mask & batch.dimension_mask
    if head_type == "regression":
        target = normalize_action(batch.actions, train_stats)
        pred = model.regress(cond)
        loss = masked_l1(pred, target, valid)
    elif head_type == "diffusion":
        noisy, time, noise = diffusion_training_pair(batch)
        pred = model.predict_noise(cond, noisy, time)
        loss = masked_mse(pred, noise, valid)
    elif head_type == "flow":
        point, time, velocity = flow_training_pair(batch)
        pred = model.velocity(cond, point, time)
        loss = masked_mse(pred, velocity, valid)
    optimizer.zero_grad(set_to_none=True)
    loss.backward()
    clip_grad_norm_(model.parameters(), max_norm=1.0)
    optimizer.step()
    if step % eval_interval == 0:
        model.eval()
        with torch.no_grad():
            evaluate_validation_rollouts(fixed_initial_states)
        save_checkpoint_and_training_state()
        model.train()
```

混合精度训练时，还要按框架启用 autocast、梯度缩放和缩放前后的裁剪顺序；上述接口省略了这些设备细节。归一化统计和 split manifest 必须固定。测试 rollout 不应在训练循环中用于挑 checkpoint。

第四部分

数据与证据：让闭环表现可以检验

第十四章　数据工程、schema 与无泄漏划分

学习算法消费张量，实验结论却依赖张量背后的物理语义和采样规则。一个缺少单位、时间和来源的数据集，即使可以成功加载，也未必能用于可复现的机器人学习。

14.1　数据规模先还原到物理时间

frame、step、episode、小时和任务数衡量不同对象。若 N_{steps} 个真实控制步以固定频率 f 采集，则

$$\text{hours} = \frac{N_{\text{steps}}}{3600 f}.$$

例如 10^7 个控制步在 30 Hz 下约为 92.6 小时；若这 10^7 是三相机图像总数，或含重采样副本，就不能作同样换算。非均匀采样应按时间戳累计有效时长。

同时报告 episode 长度分布、任务/场景/本体/操作者数量、成功/失败/接管比例与重复率。十万条相似短轨迹可能比少量长轨迹包含更多独立初态，也可能只是同一路径的重复；单一总量无法回答覆盖问题。

14.2　多源数据的对齐强度与用途

真实机器人、手持接口、人类视频、仿真和通用图文提供不同信息，数据金字塔是一种组织视角 [51]，不是规模与价值严格单调的定律。仿真可以有精确动作标签却缺少真实接触；人类视频有丰富行为却缺少机器人控制对应；图文可提供对象语义，却通常不直接监督低层动作。



图 14.1　多源数据通过表示与监督接口连接；每一类都需要独立的质量和覆盖证据。

遥操作有直接动作但有操作者风格；牵引示教受硬件支持限制；脚本或规划器便于重复但覆盖可能窄；自主 rollout 接近部署分布却需要复位和安全设施；人工接管集中在困难状态，但不是随机采样。行为来源与采样机制都应写进数据，而非只保存成功视频。

14.3　最小 schema：结构可检查，物理语义可追踪

每个 episode 保存任务、场景、本体、操作者、来源、许可、版本和划分组 ID；每个观测保存自己的时间戳、模态有效性与标定引用；每个动作分别保存建议命令、约束修改后/最终执行命令及执行时刻。T 个动作对应 $T+1$ 个状态/观测记录，最后的 s_T, o_T 不再有下一动作。

下面是一个可由 JSON 解析器读取的最小教学记录：只有本体观测，一个零位姿增量动作，

最后因示例时域结束而截断；它不是机器人实测。x^{prop} 的前七维是关节角，末维是夹爪宽度；动作前三维是平移增量（m），随后三维是旋转向量（rad），末维为绝对夹爪开度比例。此处的完整位姿增量约定在基座坐标左乘。实际系统还需给工具中心点、内外参与固件等外部 manifest。

Listing 14.1 最小 episode JSON；保存为 episode_example.json

```
{
  "schema_version": "1.0",
  "source_kind": "synthetic_teaching",
  "episode_id": "example-001",
  "task_id": "hold", "scene_id": "scene-A",
  "robot_id": "arm-A", "operator_id": "script",
  "group_id": "session-A", "license": "CC0-1.0",
  "control_dt_s": 0.05,
  "prop_layout": ["q_rad[7]", "gripper_width_m[1]"],
  "action_spec": {
    "mode": "left_se3_delta_with_absolute_gripper",
    "frame": "base",
    "units": ["m", "m", "m", "rad", "rad", "rad", "ratio"],
    "gripper_open_value": 1
  },
  "calibration_ref": null,
  "steps": [
    {
      "t_obs_ns": 1000000000, "rgb_ref": null,
      "x_prop": [0, 0, 0, 0, 0, 0, 0, 0.04],
      "action": {
        "policy": [0, 0, 0, 0, 0, 0, 1],
        "executed": [0, 0, 0, 0, 0, 0, 1],
        "source": "script",
        "t_send_ns": 1005000000,
        "t_apply_ns": 1010000000
      },
      "terminated": false, "truncated": false
    },
    {
      "t_obs_ns": 1050000000, "rgb_ref": null,
      "x_prop": [0, 0, 0, 0, 0, 0, 0, 0.04],
      "action": null,
      "terminated": false, "truncated": true
    }
  ],
  "outcome": {"task_success": false, "reason": "example_horizon"}
}
```

下面的完整标准库脚本可用 `python3 validate_episode.py episode_example.json` 运行。它检查结构、有限数值、时间顺序与末步语义；它不验证图像文件、真实时钟误差、碰撞或动力学，因此通过结构校验不能称为物理可执行证明。

Listing 14.2 最小结构校验器；保存为 validate_episode.py

```
import json
import math
import sys

def require(ok, message):
    if not ok:
        raise ValueError(message)

def finite_vector(value, size):
    return (isinstance(value, list) and len(value) == size
            and all(type(x) in (int, float) and math.isfinite(x)
                    for x in value))

e = json.load(open(sys.argv[1], encoding="utf-8"))
require(e["schema_version"] == "1.0", "schema version")
dt = e["control_dt_s"]
require(type(dt) in (int, float) and math.isfinite(dt)
        and dt > 0, "control period")
require(e["prop_layout"] == ["q_rad[7]", "gripper_width_m[1]"],
        "proprio layout")
require(bool(e["license"]), "license missing")
for key in ("episode_id", "task_id", "scene_id",
            "robot_id", "operator_id", "group_id"):
    require(bool(e[key]), "missing ID: " + key)

spec = e["action_spec"]
require(spec["frame"] == "base", "unexpected action frame")
require(spec["mode"] ==
        "left_se3_delta_with_absolute_gripper", "action mode")
require(spec["units"] ==
        ["m", "m", "m", "rad", "rad", "rad", "ratio"], "units")
steps = e["steps"]
require(len(steps) >= 2, "need action and terminal observation")
previous = None
for i, row in enumerate(steps):
    stamp = row["t_obs_ns"]
    require(type(stamp) is int, "timestamp must be integer ns")
    require(previous is None or stamp > previous, "time order")
    previous = stamp
    require(type(row["terminated"]) is bool and
            type(row["truncated"]) is bool, "end flag type")
    require(finite_vector(row["x_prop"], 8), "proprio vector")
    if row["rgb_ref"] is not None:
        require(bool(e["calibration_ref"]), "calibration missing")
    if i == len(steps) - 1:
        require(row["action"] is None, "terminal action")
        require(row["terminated"] or row["truncated"], "end flag")
    else:
```

```
        require(not row["terminated"] and not row["truncated"],
               "early end flag")
        a = row["action"]
        for name in ("policy", "executed"):
           require(finite_vector(a[name], 7), "action vector")
           require(0 <= a[name][-1] <= 1, "gripper range")
        require(type(a["t_send_ns"]) is int and
               type(a["t_apply_ns"]) is int, "command time type")
        require(stamp <= a["t_send_ns"] <= a["t_apply_ns"],
               "command timestamp order")
        require(a["t_apply_ns"] < steps[i + 1]["t_obs_ns"],
               "action after next observation")
print("valid structural teaching episode")
```

这个最小记录采用一个动作在下一观测前应用的同步约定。异步队列应增加计划 ID、目标执行时刻和每传感器时戳，采用与队列一致的校验规则；不能沿用本例断言而误拒合法延迟数据。发布数据还应记录软件/控制器/固件版本、相机内外参、深度尺度、关节限位、力和安全事件、接管起止、恢复标签、合成来源与原文件校验值。未知值用显式缺失及有效性掩码，不能与数值零混同。

RLDS 等格式统一 episode/step 存储，但存储格式不会自动统一动作坐标和控制意义。Open X-Embodiment 展示了跨数据源标准化与适配的价值 [23]；第 十一 章的动作契约仍须单独执行。

14.4 代表资源：按用途而非单一规模排列

表 14.1 数据与任务资源的接口重点；规模和可用字段应以所用 release 为准

资源	主要信息	使用时应检查的问题
RoboNet	多机器人视频与动作	多本体动作、相机和目标域适配口径，不能把迁移收益当作固定倍数 [52]
BridgeData V2	跨场景操作，语言或目标条件	任务/场景覆盖、低成本机械臂接口与语言标注版本 [53]
Open X-Embodiment	聚合多机构、多本体轨迹	各组成数据的许可、控制频率、动作映射与混合权重 [23]
DROID	分布式真实场景，多视角与标定	建筑/场景分组、相机外参与控制器版本，避免房间泄漏 [54]
RH20T	视觉、力、音频与本体信息	各传感器同步、缺失字段、配对关系和接触任务质量 [12]
RoboMIND	多本体轨迹、失败与任务元数据	质量标签、失败原因、真实/数字孪生来源与版本 [39]
MimicGen	源示范扩增的任务轨迹	源与生成副本同组、物理执行验证与成功筛选 [50]
CALVIN / LIBERO	带示范的语言条件任务	任务套件、初态、训练设置与聚合协议；详见下一章 [55, 56]
RoboCasa	厨房场景与组合任务	资产/任务版本、原子与长程任务差异、合成数据来源 [49]

14.5 按主张设计 group split

先决定要证明哪种泛化，再划分数据，不能训练完后把容易的测试变化称作“未见”。至少保留训练、调参验证和最终测试三部分；checkpoint、超参数和阈值只由训练/验证决定。

1. **共同底线：**同一 episode、相邻切片、源示范与生成/增强副本不能跨 split。共享采集 session、对象摆放和几乎重复视频的轨迹，优先按组一起划分；用 ID、时间和近重复检查补查。
2. **场景泛化：**以房间或场景 ID 为组，测试场景完全留出；任务和本体可固定。日期和操作者可以另作分层，避免把同时变化的因素都归因给背景。
3. **任务泛化：**按目标谓词、技能组合或任务模板分组，同义语言改写与同一目标的近重复不能跨组。区分新组合、新单技能和新对象语义。
4. **本体泛化：**完整留出一个机器人或本体族，明确是否允许少量目标数据、目标归一化统计和重标定；一旦使用，应单列适配设置。

对象实例泛化需要隔离实例，类别泛化还要隔离类别。存在多个泄漏关系时，可把 episode 作为图节点，共享任一禁止跨组的来源就连边，按连通分量划分；若因此无法同时平衡所有任务，应公开该限制，不能悄悄打散。

导出固定 split manifest，包含原文件校验值、分组键、随机种子和划分脚本版本。动作统计、词表、缺失值填补和数据筛选规则从训练集学习后冻结。报告每个 split 的任务、场景、本体和 episode 数量，并检查交集确实符合主张。

14.6 质量、失败数据与发布责任

质量包括动作平滑、任务进度、碰撞、接管、时间、行为一致性与是否可恢复。失败状态可以提供覆盖，失败动作是否作为正例取决于目标；条件化模仿、价值学习和恢复训练可以使用同一记录的不同字段。

家庭和人体视频可能含面孔、声音、住址与敏感物品。采集同意、许可、去标识化和撤回机制应与数据来源一起维护。第三方论文可公开下载，不代表其全部数据和素材具有相同再分发许可。

练习与提示

同一源示范扩增为 50 个视角和 20 条仿真轨迹。按帧随机划分得到低损失，能否证明新场景泛化？给出分组键、统计估计位置和近重复检查。提示：共享来源先连组，测试场景另行留出。

第十五章　仿真工具与不可混同的任务协议

仿真器提供动力学、渲染与传感器；任务环境再规定初态、目标和终止；基准协议还规定数据划分、动作接口、测试次数与聚合指标。论文使用同一软件包，并不足以证明结果可直接比较。

15.1　按科学问题选择工具

表 15.1　常见仿真生态的用途与需要核查的条件

生态	常见用途	版本与实现注意事项
MuJoCo / robosuite	刚体操作、控制、RoboCasa 等任务	接触参数、控制器和资产需要单独配置；视觉任务仍有渲染成本
SAPIEN / ManiSkill	操作、几何输入、并行学习	ManiSkill 各代物理/并行接口不同，不能把后续特性都归给 ManiSkill2
Isaac Sim / Lab	传感器、数字孪生、人形与机器人训练	引擎、驱动、资产和硬件依赖较重，记录完整版本
Habitat	导航与移动操作	场景资产和任务接口影响可见信息，精细接触需另行验证
AI2-THOR / ProcTHOR	家庭交互、语言与任务规划	离散动作和对象状态抽象与真实控制差异较大
OmniGibson / BEHAVIOR	日常活动与丰富对象状态	资源、对象模型和状态谓词影响可复现性
PyBullet / Gazebo	教学、控制原型与系统集成	模型、插件和任务库并非统一基准

吞吐、接触精度和视觉真实感均依赖任务、引擎、版本与硬件。选择工具应先问需要控制什么变量，而不是先排一个“最真实”名次。若研究动作头，优先固定短程任务与控制器；若研究长程规划，需要可检查的技能和恢复；若主张 Sim2Real，最终仍需独立真实验证。

15.2　导航与家庭交互提供什么证据

Habitat 从具身导航扩展到移动操作和家庭重排 [57, 58]；ProcTHOR 用程序化房屋增加场景变化 [59]；ALFRED 结合语言、导航与对象状态变化 [60]；BEHAVIOR-1K 组织丰富的日常活动与对象状态 [61]。它们能检验语义、记忆和规划的某些方面，但传感器、可用真值、动作离散化与接触抽象必须说明。

例如，允许查询对象真值位置的规划器与只看 RGB 的策略，不是在同一信息条件下竞争；用符号“打开抽屉”动作也不能直接证明低层机器人能完成真实把手接触。

15.3 RLBench：task、variation 和 episode 三层

RLBench 原论文组织了 100 个手工任务，并区分三层 [62]：

Task 动作结构与成功条件，如堆叠或开盖；

Variation 同一任务中的目标对象、颜色或语言条件等变化；

Episode 给定 variation 后的具体初始位姿和一次执行。

因此“未见 episode”可能只意味着新摆放，不能自动称为未见技能。评测子集、variation 分组和初态种子应公开；后续常用任务子集也不能称作整个 RLBench 的统一协议。

环境支持多类关节或末端动作模式。直接关节控制、末端位姿加 IK、带规划器的末端到达 wrapper，会给策略提供不同能力。若一个动作触发内部规划路径，就不能把策略调用次数直接当低层控制步数。相机集合、深度/点云、关键帧抽取、碰撞检查、控制器和 task-specific success detector 都属于协议。

15.4 CALVIN：连续子任务链与环境泛化

CALVIN 定义了 34 个操作子任务，并用连续五个子任务的链评估长程语言条件控制 [55]。它提供静态与手眼 RGB-D、本体以及视觉触觉输入的接口；实际方法使用哪些模态，必须以配置为准，不能因环境支持触觉就默认每个模型都用了触觉。

常见设置用环境 A/B/C 的数据训练、在 D 测试；论文也研究其他训练环境组合。每条链按顺序执行，链中不在每个子任务后重置物体状态，失败会影响后续进度。需要报告测试链列表、每子任务最大步数、语言条件切换、成功 oracle、动作模式和初态数。

设第 i 条链连续完成 $L_i \in \{0, \dots, 5\}$ 个子任务，可报告平均链长

$$\overline{L} = \frac{1}{N}\sum_{i=1}^{N} L_i = \sum_{j=1}^{5} \widehat{\Pr}(L \geq j).$$

它不是单任务成功率，也不能把 $\overline{L}/5$ 直接当作“五任务整链成功率”；后者是 $\widehat{\Pr}(L = 5)$。只有在额外且通常不成立的独立同分布子任务假设下，才可能从单步概率推算整链概率。

15.5 LIBERO：原始终身学习与后续多任务协议

LIBERO 原始资源包含 Spatial、Object、Goal 各 10 个任务以及 LIBERO-100 共 100 个任务，总计 130 个 [56]。前三组分别控制空间关系、对象概念或任务目标等变化；LIBERO-100 含更混合的知识要求，原论文将其划为 LIBERO-90 与十个长程任务。

原论文研究顺序学习的知识迁移与遗忘，需记录任务顺序、每阶段训练预算、历史任务表现及相应曲线指标。后续 VLA 常报告 Spatial/Object/Goal/Long 四组各十个任务的多任务或微调成功率 [15]；这不是对全部 130 任务的同一终身学习实验。报告“四套件平均”时必须列出套件和具体版本。

物理引擎、机器人控制器、动作 wrapper、夹爪符号、初始状态文件、相机、超时、稳定等待、每任务试验数与成功谓词都会影响数值。选择 checkpoint 的验证任务、训练示范数量和数据清洗也必须一致。ManiSkill2、RoboCasa 等其他操作资源同样需要这些说明 [49, 63]。

15.6 为什么三个基准的百分比不能横向排名

表 15.2 协议差异决定指标含义；表中为需具体化的接口，不是统一排名

基准	测试单位	动作与执行	典型聚合与主要风险
RLBench	task/variation 下的 episode	关节或末端；可能使用 IK/规划 wrapper	选定任务成功率；任务子集与隐藏规划能力改变问题
CALVIN	环境中的五子任务链	连续动作与子任务切换，无中途重置	前缀成功比例、平均链长；不是独立短任务百分比
LIBERO	明确套件/初态的任务 rollout	指定机器人控制与动作 wrapper	多任务成功率或终身迁移曲线；两类设置不能混报

成功率依赖任务难度、可用信息、控制器、时域和成功定义。即使两个数字都为 80%，它们可能分别表示有规划器的末端任务、五任务链的第一步、或特定套件的宏平均。跨基准比较应讨论能力覆盖和失败机制，不能把数字相减得出普遍优劣。

15.7 可复现实验配置应落到文件

协议清单应具体到文件；下例仅展示字段，须按实际环境生成并冻结相应文件。

Listing 15.1 任务协议 manifest 的字段示例

```
protocol = {
    "environment": "name@commit", "task_suite": "task-list.json",
    "train_split": "train.json", "validation_split": "validation.json",
    "test_initial_states": "fixed-state-list-and-checksum",
    "observation": ["camera_names", "proprio_fields"],
    "action": ["mode", "frame", "units", "controller"],
    "control_dt_s": 0.05, "H_o": 2, "H_p": 8, "H_e": 2,
    "episode_limit": "low-level-steps-or-physical-seconds",
    "success_rule": "predicate-and-version",
    "intervention_rule": "autonomous-and-assisted-accounting",
    "checkpoint_rule": "chosen-on-validation-only",
    "evaluation_seeds": [0, 1, 2]
}
```

0.05 秒与这些时域只是字段示例，不是三个基准的统一默认值。实际复现还应保存软件依赖、资产、策略与归一化统计校验值、逐任务结果、失败轨迹及运行命令。

练习与提示

CALVIN 平均链长均为 4：A 每条完成四步；B 有 80% 全成、其余首步失败。整链成功率相同吗？能否与 LIBERO 的 80% 直接比较？提示：列出 L 的分布，区分均值与尾部概率。

第十六章　闭环评测、统计区间与失败恢复

评测要回答两件事：系统是否完成任务，证据支持多大范围的结论。先固定成功与接管口径，再汇总统计；先给出任务和接口，再解释百分比。离线误差、论文自报和本次实验必须分开。

16.1　成功率的 Wilson 区间

若固定协议下 $n > 0$ 次试验可近似视为独立、同一成功概率的伯努利试验，成功数为 k，则 $\hat{p} = k/n$。Wald 标准误差 $\sqrt{\hat{p}(1-\hat{p})/n}$ 在小样本或 $\hat{p}$ 接近 0/1 时表现差，尤其会错误地产生零宽区间。Wilson 分数区间更适合此类报告 [64]。令正态分位数 $z = 1.96$，近似 95% 区间的中心与半宽为

$$c = \frac{\hat{p} + z^2/(2n)}{1 + z^2/n}, \tag{16.1}$$

$$w = \frac{z}{1 + z^2/n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}, \tag{16.2}$$

$$[p_-, p_+] = [c - w, c + w]. \tag{16.3}$$

“95%”描述重复采样时区间方法的覆盖性质，不表示一次区间中参数以 95% 概率随机取值，也不消除任务定义或选择偏差。区间的分子、分母和置信水平必须公开。

手算 $k = 6, n = 10$ 得到约 $[31.27\%, 83.18\%]$；0/10 的上限约为 27.75%，10/10 的下限约为 72.25%。这些是独立的统计算例，不是下文二十条记录的汇总，也不是模型成绩。零次成功不能证明成功概率为零，十次全成功也不能证明必然成功。

Listing 16.1　只依赖标准库的 Wilson 区间函数

```python
import math

def wilson(k, n, z=1.96):
    if type(k) is not int or type(n) is not int:
        raise ValueError("k and n must be integer counts")
    if (n <= 0 or not (0 <= k <= n)
            or not math.isfinite(z) or z <= 0):
        raise ValueError("need n > 0, 0 <= k <= n, z > 0")
    p = k / n
    denom = 1 + z * z / n
    centre = (p + z * z / (2 * n)) / denom
    half = z / denom * math.sqrt(
        p * (1 - p) / n + z * z / (4 * n * n))
    return max(0.0, centre - half), min(1.0, centre + half)
```

多个任务的成功概率通常不同。应给逐任务次数与区间，并声明宏平均（任务等权）或微平均（试验等权）。不能把宏平均直接代入总 k/n 的同概率假设；同一场景、机器人或初态下的重复试验也可能相关。需要时按任务、场景或独立实验批次做分层/聚类重采样，并说明推断对象是固定任务集还是新任务。配对初态的 A/B 试验应利用配对结构，不把相关试验伪装成独立样本。

16.2 六轴泛化：留出什么，固定什么

1. **视觉：**背景、光照、遮挡、相机与纹理；
2. **对象：**实例、类别、尺寸、材质与形状；
3. **几何与运动：**初始位姿、目标、轨迹与障碍；
4. **语义：**未见指令、属性、组合与词汇；
5. **任务：**新技能、长程组合、顺序与条件分支；
6. **本体：**机器人、夹爪、自由度、控制模式与频率。

每个轴给出训练/测试 ID、划分规则与其他不变条件。只换背景不能证明技能泛化；相同实例换颜色不能证明新对象类别。多个因素同时变化可以测整体分布偏移，但不能据此识别单因素原因。无泄漏划分见第 十四 章。

16.3 同一流程中的任务、时延、干预和安全指标

表 16.1 建议同时报告的指标及分母

指标	口径	必须说明的条件
自主成功率	无人工接管且满足任务成功 / 全部纳入试次	允许自主重规划和自主恢复的范围
最终完成率	含规定辅助后的成功 / 全部纳入试次	辅助者、时间预算与接管规则
干预	接管 episode 比例、事件数/episode、事件数/小时	一次事件起止、总运行时长、重复接管
恢复	恢复成功 / 明确定义的机会或尝试	自主/人工分开；不尝试的机会是否纳入
时序与资源	实际 Hz、端到端时延及观测年龄分位数	测量位置、设备、队列、网络与预处理
安全	碰撞、越界、峰值力、shield 修改与停止	事件阈值、暴露时长、是否排除试次
任务代价	完成时间、路径、能耗与动作变化	失败如何计时，平滑与 jerk 的单位/差分规则

自主恢复后完成可以计自主成功，但另标恢复。人工接管后完成属于辅助成功，不能回填为原策略成功；接管后仍失败既计接管也计失败。若在试次开始前就有规定的共享控制，应单独命名系统和协议。

恢复率的分母尤其容易混淆：“尝试过的恢复成功率”会排除未识别或未尝试的异常，不等于所有异常的挽救率。建议同时记录异常机会、是否检测、是否尝试、是否成功以及最终任务结果；一次 episode 多次恢复时使用事件 ID。

时延至少报告曝光/采样至实际执行的 p50、p95 和最大值，并区分纯推理时间、队列等待和观测年龄。成功率相同而 p95 时延不同的系统，在接触任务中可能有不同风险。干预每小时需要真实累计运行时间，仅有 episode 数量无法计算。

16.4 二十条合成记录怎样汇总

本项目保存了二十条**合成教学评测记录**，其中的结果、时延和离线误差均为人为构造，不来自机器人或本次训练模型。原始 CSV、统计输出与计算命令见第 十八 章及配套文件 `output/experiments/evaluation_records.csv`。

表 16.2 合成教学日志的口径演示，不能解释为实测模型性能

项目	统计	解释
自主成功	11/20 = 55%	包含两次自主恢复；Wilson 约 [34.21%, 74.18%]
含辅助最终完成	14/20 = 70%	包含三次人工后完成；Wilson 约 [48.10%, 85.45%]
最终失败	6/20	与是否接管、是否曾恢复分别记录
接管 episode / 事件	4/20；共 5 次	一个 episode 可有多次接管，二者分母不同
恢复尝试 / 成功	5 次 / 3 次	每次位于不同 episode；成功中两次自主、一次人工
出现主异常的 episode	11/20	包含后来恢复或人工完成者，不等于最终失败数

这组记录的主异常互斥计数为：接触 3、时序 2、感知 2、动作 2、规划 1、安全 1，共 11。这些是“出现异常”的分类，不能拿它们当作六个最终失败的分类。若要分析最终失败，应先筛选最终失败行，再按其分类统计。现实中主因往往是审阅判断，应另保留次因和证据，避免把多重原因强行变成确定因果结论。

日志未记录总运行时长，因此不能计算接管每小时；其每条时延是教学数值，不能支持硬件延迟结论。统计程序可对这些值求分位数，但计算正确不改变数据来源性质。

16.5 为什么离线动作误差不够

离线误差在固定数据分布上衡量预测与标签距离；闭环性能还依赖学习者访问分布、动作多解、控制器、延迟、终止和恢复。一个略偏的动作可能把系统带到未覆盖区域；两个不同但都可行的动作又可能有较大标签误差。

教学日志中，S15 的离线 MSE 为 0.001 却失败，S07 的 MSE 为 0.002 且自主成功，S14 同样为 0.002 但依赖人工完成。这些人为安排只演示“单一误差数值不决定结果”的逻辑，不构成经验相关性研究。第 十八 章的独立轻量仿真消融才是本次实际训练与 rollout 证据，范围也仅限其环境。

评测动作误差仍有价值：它可以发现单位错误、标签错位和某些过拟合。应将其作为诊断之一，与固定初态闭环试验共同解释。

16.6 公平比较、故障与失败分析

比较方法时固定训练/目标数据、输入、动作控制、频率、H_o, H_p, H_e、初态、随机种子、最大步数和 checkpoint 选择规则。报告参数、显存、训练计算及实际端到端性能；不同硬件或不同控制器的结果可分别报告，但不能把差异全归给算法。

预先定义硬件故障、环境加载失败、网络中断与安全停止是否计入主分母。与真实部署相关的系统故障通常应计入系统指标；若另给“排除故障”的算法指标，需要同时公开全部试次及排除原因，不能事后只删不利记录。

失败可分感知/指代、计划、可达性/IK、动作、接触、时序、终止、恢复、安全与系统故障。保留主因、次因、发生阶段、可恢复性和视频/日志证据。压缩研究还应检查失败类型是否变化：空间精度受损、微小接触动作消失和推理变快分别需要不同解释。

截至 2026-10-22，公开与非公开系统覆盖的任务、数据和协议仍不同；本书不设跨全部具身能力的统一排名。任何性能表述都应带上任务、数据、日期、协议和“论文报告/独立复现/本次实测”的来源标签。

练习与提示

合成日志中有四个 episode 发生接管，共五次事件。分别计算接管 episode 比例和平均每 episode 接管事件数；为什么不能得到每小时干预数？若所有出现异常但未尝试恢复的试次也纳入分母，恢复率的解释会怎样改变？

提示：先写分子与分母的单位，再给指标命名，不能只报“恢复率 60%”。

本章要点

可信评测由固定协议、分层测试、统计不确定性、时序资源与事件日志共同组成。自主、辅助、恢复和异常是不同字段；把它们放在同一流程中，才能解释系统何时成功、为何失败和结论适用于哪里。

第十七章　运行时安全、控制约束与保证边界

平均成功率高，不表示碰撞概率低；模型报告不确定，也不表示系统会及时停下。安全论断必须指出约束对象、系统模型、执行机制和证据。本章区分三类常被混用的概念：运行时过滤、控制约束与有假设的数学保证。

17.1　风险来自整个感知与执行链

VLA 同时面临视觉误导、语言指令冲突、错误指代、训练后门、远程服务故障和动作生成错误。长程系统还可能受记忆污染、过期计划和错误成功判断影响。相关综述组织了具身模型的威胁与评测问题 [65]；具体风险仍取决于机器人、任务、控制和人员环境。

安全层不能只检查自然语言。一个语义上合法的“把杯子放回去”，也可能因杯子被移动、标定错误、路径有人或夹爪未闭合而危险。对目标、技能、动作和硬件分别设置可检查条件，并记录哪一层做了修改。

17.2　分层机制与各自的证据

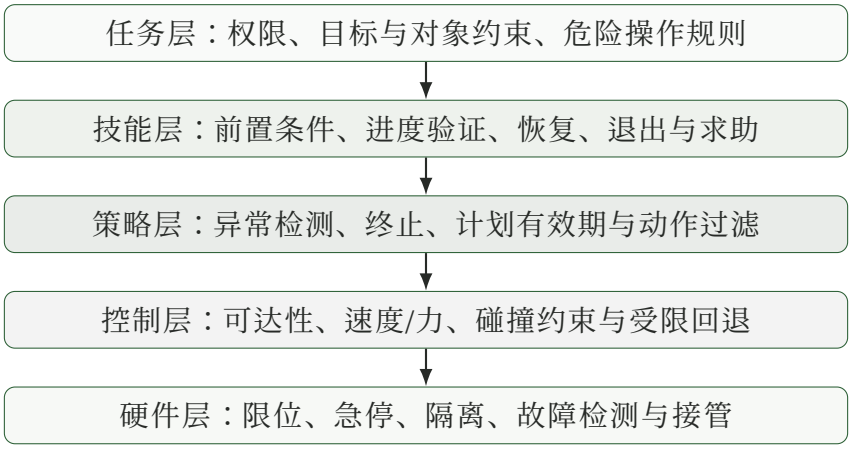


图 17.1　安全机制需跨层配合；每一层都可能受模型、传感器、时序或硬件故障限制。

任务和技能层的契约可限制允许的目标与调用顺序；策略层可拒绝明显异常动作；控制层把高层命令转换为受限运动；硬件层降低失控后果。冗余机制有价值，但不能因“有急停”就声称边界绝不会突破。检测、通信、制动距离、力学惯性和故障独立性都需要测试。

17.3　不确定性要经过风险校准

Ensemble 分歧、MC dropout、动作熵、flow 采样方差、视觉距离、世界模型预测误差和 critic 分歧可作为异常线索，但它们的量纲和含义不同。多峰动作分布可能只表示“左右绕行都可行”，并非危险；模型也可能对错误动作很自信。

应在独立验证数据上检验评分与失败的关系，固定阈值后再测试。若将高风险部分交给回

表 17.1 三类安全表述的对象和限度

机制	典型实现	可检验的性质	不能直接推出的结论
运行时 shield	规则、动作限幅、碰撞筛查、回退	哪些建议被改写，触发和漏检率	长期任务安全或所有未知场景安全
控制约束	动力学 QP、速度/力限制、CBF	给定模型下约束可行与跟踪误差	感知、模型和时延任意失准仍安全
形式化保证	不变量证明、可达性、契约或状态机验证	明确抽象、初态与假设下的性质	抽象外硬件与现实环境自动满足假设

退控制器，可画风险 – 覆盖率曲线：覆盖率是自主执行比例，风险是这部分中的失败或安全事件率。应同时报告被拒绝部分和回退的表现，避免只保留容易样本制造高成功率。

风险响应可以是降低速度、增加观测、缩短 H_e、重新规划或求助；停止也需判断是否安全，例如悬持物体和接触施力时不能简单断电。评分、决策和实际动作间的时延必须进入评测。

17.4 一个可推导的 CBF 条件

考虑简化的连续时间控制仿射系统

$$\dot{s} = f(s) + G(s)u, \qquad \mathcal{C} = \{s : h(s) \geq 0\}.$$

这里 s 是控制模型的真实状态，不是任意视觉 embedding，u 是该动力学模型中的控制输入；VLA 输出的位置目标或 chunk 需要先转换到这个接口。假设 h 可微，动力学及反馈满足解存在唯一和持续存在的条件，相关状态可获得，输入可行，且无未建模扰动或延迟。

对于相对阶为一、边界适当正则的约束，可选择 $\kappa > 0$ 并要求

$$\nabla h(s)^{\top}[f(s) + G(s)u] + \kappa h(s) \geq 0. \tag{17.1}$$

这是常用控制屏障函数（control barrier function，CBF）条件的一种简化线性形式 [66]。沿满足条件的连续轨迹，

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(e^{\kappa t}h(s(t))\big) = e^{\kappa t}(\dot{h} + \kappa h) \geq 0.$$

因此初始 $h(s(0)) \geq 0$ 时，$h(s(t)) \geq e^{-\kappa t}h(s(0)) \geq 0$。这个结论依赖上述条件在整个执行期间成立；它没有证明真实视觉机器人已满足这些假设。

可在低层求解

$$u^{*} = \arg\min_{u \in \mathcal{U}} (u - u^{\text{policy}})^{\top} W (u - u^{\text{policy}}) \tag{17.2}$$

$$\text{s.t. } \nabla h^{\top}(f + Gu) + \kappa h \geq 0, \tag{17.3}$$

其中 W 为正定矩阵。在控制仿射、约束对 u 线性且 $\mathcal{U}$ 为适当凸集等条件下，它可写成小型凸二次规划。高相对阶、接触切换、多个冲突约束和饱和输入需要额外处理；若 QP 无解，不能用未说明的裁剪继续声称满足条件。应记录不可行事件并进入已测试的回退。

17.5 采样、延迟和估计误差会改变保证

一个一维手算说明只在采样点检查的局限。令安全距离为 $x \geq 0$、动力学 $\dot{x} = u$、$h(x) = x$。连续反馈 $u = -\kappa x$ 使 $x(t) = x(0)e^{-\kappa t}$，不会越过零。若只在 $t = 0$ 求一次动作并保持 Δt，则

$$x(\Delta t) = x(0)(1 - \kappa \Delta t).$$

例如 $\kappa = 20\,\mathrm{s}^{-1}$、$\Delta t = 0.1\,\mathrm{s}$ 时，初始 $x = 0.1\,\mathrm{m}$ 会变为 $-0.1\,\mathrm{m}$。初始时刻满足式 (17.1)，保持输入期间却不再持续满足，因此不违背上面的连续时间推导。

真实控制需采用与采样周期匹配的离散时间条件、可验证的鲁棒裕度或采样间轨迹检查，并考虑计算/通信延迟、估计误差和制动距离。模型未知时的规则 shield 可以降低部分风险，但不能自动获得基于精确动力学的 CBF 保证。若放松屏障约束加入 slack，应明确保证已转为软约束。

17.6 形式化接口与可测试边界

可以对技能有限状态机、调用权限、抽象工作空间和终止逻辑验证明确性质。例如“只有确认夹爪已持物后才移动到交付区”可成为技能前置条件；但传感器是否正确识别持物，仍是该保证的外部假设。

本书不提供完整 VLA 与现实环境的端到端形式证明。更可操作的做法是列出：受证明的模块、抽象模型、初态集合、传感器/控制误差界、时序假设、故障处理及未覆盖情形。形式证明、仿真测试和真机测试提供不同证据，应各自标注，不相互替代。

日志应保存原策略建议、shield 或 QP 修改后的动作、最终执行动作、约束余量、求解状态、时间戳与触发原因。训练时若只存原建议，下一状态却来自修改后动作，就会产生错误转移标签。

17.7 故障注入与安全评测

在受控条件下测试遮挡、错误相机、传感器冻结、网络断连/抖动、目标被移动、不可达目标、夹爪空抓、语言与视觉冲突和重复失败。对每种注入记录是否检测、检测延迟、错误继续执行、回退成功、人工介入与任务后果，并统计正常场景的误触发成本。

安全停止通常应计入系统试次与事件分母；若停止阻止了危险，它可以是安全机制的成功，同时仍是任务未完成。第 十六 章的多字段记录正用于保留这种区别。

练习与提示

若关节速度被限制在允许范围内，是否足以保证手持物体不会撞到人？用一个具体轨迹说明还缺哪些状态、几何、时延和力学条件。再解释为什么给安全 QP 加 slack 后不能沿用无松弛条件的不变性结论。

提示：逐时刻的单变量限幅与整条系统轨迹的安全集合是不同约束。

本章要点

规则过滤、控制约束与形式证明各有明确对象。安全结论必须同时携带模型、状态、输入可行性、采样和故障假设；日志与故障注入检查这些假设在执行中是否成立，而不是用平均成功率替代安全证据。

第十八章　最小实验：动作时域、扰动与评测口径

本章交付三种不同证据：第 2.5 节的二维两连杆手算用于验证接口；本章的轻量点质量仿真是实际运行的行为克隆消融；人工编写的评测记录用于演示统计。点质量仿真没有机械臂运动学、视觉、接触或安全认证，不能替代机器人基准。

18.1　本次运行：行为克隆与动作块长度

本节是本书配套脚本的实际运行结果，环境为二维点质量教学仿真，**不是真实机器人或公开基准复现**。位置、速度和加速度的单位分别为 m、m/s 与 m/s^2。策略输入为目标相对位置与速度，动作是二维加速度；以 10 Hz 更新

$$v_{t+1} = \mathrm{clip}(v_t + 0.1a_t, -1, 1), \qquad p_{t+1} = p_t + 0.1v_{t+1},$$

其中每轴加速度先裁剪到 $[-2, 2]$。专家采用带饱和的 PD 控制 $a_t = \mathrm{clip}(6(g - p_t) - 4v_t, -2, 2)$。

按 episode 固定为 96 条训练、24 条验证与 64 条测试，每条 80 步；数据种子为 20261022。标准化统计只在训练集拟合。每个 MLP 的隐层宽度为 64，训练种子为 7、17、27，均以 batch 128、1500 次 Adam 更新、学习率 0.003 训练；同一种子各条件使用相同 minibatch 索引，取最后一次更新。没有通过测试集选模型。参数量与计算量随动作块长度变化，因此这是相同数据和更新预算的消融，并非等算量比较。

本消融取 $H = H_p = H_e$，分别设为 1、4、8；预测 H_p 步，执行全部 H_e 步后重新观测，不计推理延迟。成功要求最后连续 5 步距离不超过 0.05 m 且速度不超过 0.08 m/s。扰动条件在预先固定的第 52–65 步之一施加模长 0.8 m/s 的随机方向速度冲量。每个训练种子各评测 64 个共享初态。

表 18.1　二维教学仿真的本次实测。成功率为三个训练种子的均值；MSE 单位为 (m/s^2)^2。

$H = H_p = H_e$	参数量	首步测试 MSE	无扰动成功率/%	扰动成功率/%
1	450	0.001210	100.0	78.6
4	840	0.001469	100.0	36.5
8	1360	0.001653	100.0	7.8

所有单种子成功次数、Wilson 区间、离线损失、逐轨迹结果和训练曲线均由脚本保存。离线 MSE 在专家未扰动轨迹上计算；扰动后策略进入不同状态分布，因此离线误差不能代替闭环成功率。这一结果只适用于上述点质量、PD 专家和预算，不支持真实机器人上的性能排序；预测长度与执行长度同时变化，也不能把差异归因于其中一项。进一步消融应固定预测长度，单独改变执行长度。

18.2 合成记录：接管、恢复与成功率口径

另外构造 20 条**合成教学记录**，与上述实测严格区分。9 条直接自主成功，2 条经自主恢复成功，3 条依赖人工后完成，6 条最终失败。自主成功率为 11/20 = 55%，Wilson 95% 区间为 [34.21%, 74.18%]。若计入辅助完成，则为 14/20 = 70%，区间 [48.10%, 85.45%]；后者必须另称含辅助最终完成率。

这些区间只演示公式计算，合成计数不支持真实总体推断。实际应用要求固定协议下可视为独立的 Bernoulli 试验；同一初态反复测试或环境漂移产生相关性时，应分层报告或按簇重采样。

4 个 episode 共发生 5 次接管；5 次恢复尝试中 3 次成功，其中一次依靠人工。每个恢复 episode 在此恰好尝试一次，事件数与 episode 数因此相等。恢复率 3/5 与总体成功率的分母不同。11 条有异常的记录按首要异常互斥计数：接触 3、时序 2、感知 2、动作 2、规划 1、安全 1；它们包含恢复成功和辅助完成的记录，不是 6 条最终失败的分类。

合成的 S01 动作 MSE 为 0.020 却自主成功，S15 的 MSE 为 0.001 仍因感知异常失败。这些人为设定的数值只展示指标之间不存在逻辑蕴含，不能当成统计相关性或论文证据。

18.3 复现入口与边界

脚本位于 `manuscript/experiments/run_experiments.py`，配置为同目录 `config.json`；从本主题目录运行

```
python3 manuscript/experiments/run_experiments.py
  --output output/experiments
```

依赖 Python 3.10 以上与 NumPy。输出保存示范、episode 划分、模型权重、环境版本、完整命令、种子、逐轨迹 CSV 与报告。独立检查包括 Wilson 已知参考值与无效输入、动力学单步单位、动作块不跨 episode、成功阈值边界及两连杆正运动学参考值。

本次未运行视觉 VLA 微调、真实机械臂、接触仿真或跨本体实验。复现到机器人时，应分别控制预测长度 H_p 与执行长度 H_e，以同一数据划分、控制器和初态清单测量任务成功、接管与端到端时延；教学点质量的数值不能外推到这些设置。

18.4 从最小结果走向机器人复现

若要把这个消融迁移到 RLBench 或 LIBERO，先固定任务套件、仿真器版本、动作模式、控制周期、初态种子和成功判据；再将同一 episode 划分用于所有模型，只从训练集估计动作统计。保持 H_p 不变、只改变 H_e 可进一步分离执行反馈效应；保持 H_e 不变、只改变 H_p 则研究预测监督的影响。还应记录实际推理延迟，避免模拟中零计算等待掩盖欠载。

本次尚未安装 RLBench 或 LIBERO 的完整环境，也未采集真机、接触传感或硬件时延数据。复现可从本章脚本起步，再按第十五章的协议表选择单一任务与动作配置。进一步加入时序延迟、动作历史和图像输入时，应另开实验 ID 并保留本章基线。

第五部分

研究实践：把问题变成可证伪实验

第十九章　研究地图与前沿证据的边界

19.1 如何使用这张地图

本章依据截至 2026-10-22 已核查的随书材料组织研究线索。第六章已记录代表模型的接口与版本，这里只问：新工作改变了闭环中的哪个量，证据支持到哪一步？同一个系统可能同时包含数据共训、连续动作、世界模型和技能验证，不能把它的全部收益归因于某个名称。

19.2 从单步预测扩展到状态、计划与验证

扩大 VLA 只能改变策略的一部分。历史编码帮助估计遮挡与运动，后果预测辅助选择候选，技能契约负责有限执行，验证器决定是否恢复。这些模块对应不同失败来源，应分开记录和消融，而不是把整套系统与单一策略直接排名。VERA 将视频规划器与本体 inverse dynamics 分离 [36]，EmbodiedSkills 检查技能前置条件和执行结果 [38]；两者都提示接口可以成为研究变量。但视频预测、低层执行成功和整段长程成功是不同证据，不能互相代替。

19.3 从成功示范扩展到异质经验

机器人数据可以记录成功、失败、纠正、接管、奖励和恢复过程。异质经验首先需要 provenance、时间戳、质量/策略条件与适用的学习目标。$\pi_{0.7}$ 使用元数据条件吸收混合质量经验 [26]；共训研究比较视觉语言、密集语言、跨本体数据、人类视频与动作 token 等来源 [47]。它们支持研究数据配方，不支持“所有额外数据都有益”。数据数量、动作接口、混合比例与目标域相似度共同影响结果。失败片段可能适合训练检测或恢复，未必适合无条件行为克隆。

在线或离线后训练还改变策略访问的状态分布。LWD 使用 fleet 交互与干预经验，结合价值学习和 flow 策略提取 [44]；Z-1 在 RoboCasa 上研究 flow VLA 的组相对后训练 [45]；GRAFT 研究细粒度操作中的局部视觉监督与在线适配 [46]。这些算法的奖励、分布和动作参数化不同，不能直接把语言模型的 token 概率比套到确定性连续控制上。具体目标与边界见训练部分。

19.4 动作表示和部署预算一起设计

逐维离散、FAST 的频域 token、diffusion/flow 专家和连续回归代表不同的表达与计算折中 [15, 17, 22]。“动作头更大”与“更快连续 chunk”都可能在指定任务中改善表现，单一总参数量不足以解释收益。应固定数据与接口，分别改变专家容量、训练目标、H_p、H_e、推理步数与控制器，才有机会识别原因。

SmolVLA 的异步执行说明系统栈会影响数据时效 [14]。量化、缓存和服务端批处理的收益也不能只用每秒 token 衡量：排队、网络抖动、过期动作以及接触阶段的反馈年龄都会改变闭环。后续可在 fleet 服务与大型 VLA 上测量硬件吞吐，并记录排队、网络抖动和动作时效。

19.5 近期材料：机制与仍待验证的部分

表 19.1 随书近期材料的选择性索引，均为论文自报，未由本次实验复现。

工作与保存版本日期	本书使用的机制证据	不能据此推出
$\pi_{0.7}$，v2，2026-04-24	历史、质量/策略元数据和子目标条件	元数据一定校准，或失败动作可无条件模仿
LWD，v3，2026-08-24	fleet 经验、干预、价值与 flow 策略提取	所有机器人可按相同成本持续学习
VERA，v1，2026-05-27	视频 planner 与本体 IDM 分离	视频未来唯一决定机器人动作
Z-1，v1，2026-06-30	RoboCasa 中 flow VLA 组相对后训练	同样增益可直接外推到真机/新奖励
GRAFT，v2，2026-08-28	局部区域辅助监督与在线适配	任意任务都依赖相同区域或得到相同增益
Embodiment Gap，v1，2026-08-19	区分可复用结构与目标执行适配工作	已形成通用、唯一的部署成本标准
EmbodiedSkills，v1，2026-09-01	可执行技能契约和事件记录	强低层策略等于可靠长程记忆能力

表中版本对应文献 [1, 26, 36, 38, 44–46]。保存版本的日期与首次发表年份不同；后续复现应重新检查项目与权重版本，不能只更新“截至日期”就沿用旧性能数字。

本章要点

前沿地图的用途是找到变量与证据缺口。把“更大、更通用”的叙述拆成观测、动作、数据、价值、验证和计算条件，才能形成可比较、可证伪的研究问题。

第二十章　仍未解决的核心科学问题

20.1　跨本体共享什么，执行时还要补什么

从人类或另一机器人视频推断动作通常不唯一：相似物体轨迹可由不同抓点、关节和接触力产生。共享语义、对象关键点、末端轨迹、接触序列和 latent skill 是不同假设。可复用表示不等于免去标定、IK、控制器和数据适配 [1]。检验时应留出整个目标本体，并报告配对数据、人工标注、接口实现和目标示范成本；只报告目标成功率会隐藏适配工作。

一个可测问题是：相同目标数据预算下，共享对象/接触表示是否比共享关节向量更便于迁移？需要固定目标控制器与训练预算，分别测少样本曲线、不可达率和接口错误，而非仅看重建误差。若目标本体完全依赖专属头和大量示范，结论应收缩为预训练初始化收益，不能称为无需适配的跨本体控制。

20.2　长尾覆盖与物理反事实

透明物体、反光、柔性材料、松动夹爪、突然移动的目标与网络卡顿会产生训练分布之外的组合。更多常规成功示范不保证覆盖这些状态。novelty、critic 不确定性和模型预测误差可以作为采样信号，但可能对纹理噪声过敏，也可能对重要失败过度自信。应把采样信号的标定集与最终测试集分开，并与等预算随机采样比较。

物理反事实问的是改变动作介入后的结果。仅从“杯子在桌上”的图像不能确认从哪边推会倾倒。可在轻量仿真中随机分配推力方向或抓点，保持物体/初态分布一致，比较预测的动作排序与实际后果。只观察不同场景里的相关性不足以归因于动作机制；世界模型的预测还要接受真实或仿真交互验证。

20.3　长程信用、可更新记忆与持续学习

长任务的终点奖励难以区分早期动作贡献。子目标谓词、过程奖励和分层价值提供更密集信号，同时也可能引入错误验证与奖励投机。研究时应将低层动作错误、错误子目标、记忆过期和验证误判分开，否则一个成功率无法定位机制。状态记忆需要证据、失效与冲突处理，不能以检索相似文本替代当前环境确认。

部署后更新又会改变旧任务能力。可比较 replay、保守更新或 adapter 路由，但必须在固定的旧任务和新任务上测前向迁移、后向迁移与遗忘。更新模型、传感器或控制器后都应保留版本，先进行回归和影子运行；一个新版本的局部收益不是旧版本轨迹可直接合并的理由。

20.4　接触与灵巧操作中的充分信息

布料、线缆、食物与工具交互包含不可见接触、形变和摩擦。视觉、触觉、力与本体状态提供不同时间尺度的信息。可测问题是某个传感器在何种遮挡/接触阶段补足观测，而非“多模态是否更好”。固定数据量和策略容量，测试通道移除、时间偏移、噪声和缺失；若收益在同步被

打乱后仍存在，要排查任务 ID 或采集者等泄漏。高频传感器并不意味着高层 VLA 必须同频运行，低层控制器的带宽与稳定性同样要报告。

20.5 拒绝、求助与风险覆盖曲线

看不清、动作多解、目标不可达、计划已失败和任务已完成，需要不同处置。允许拒绝执行后，模型可通过少做难任务降低表面失败率。以下以完整任务为单位，预先固定接受自主执行的规则；执行中途接管按既定规则计入失败/辅助结果，不能事后从已接受任务删除。因此应同时定义

$$\mathrm{coverage}(\eta) = \frac{N_{\text{自主执行}}(\eta)}{N_{\text{全部任务}}}, \qquad \mathrm{risk}(\eta) = \frac{N_{\text{自主失败}}(\eta)}{N_{\text{自主执行}}(\eta)}.$$

η 是预先标定的阈值；分母为零时风险不定义。曲线之外还要报告全部任务上的自主成功率、求助、恢复成本和安全事件。平均置信分没有说明拒绝是否集中在真正困难状态，也不等于风险概率已校准。

20.6 效率误差何时会变成轨迹误差

压缩可能降低动作分辨率，也可能通过减少延迟改善反馈。权重误差、离线动作误差和闭环稳定性之间没有无条件单调关系。局部轨迹误差界通常需要动力学/策略的连续性、观测误差和延迟有界等假设；碰撞或接触切换可能破坏这些条件。因此有限任务试验应称为经验闭环表现，不能称为一般稳定性证明。量化与部署的通用实现见《智能体模型量化压缩与部署》，本书关心它们如何改变实际命令、时序与任务结果。

练习与提示

从本章选一个问题，写出介入变量、固定条件、两项互相竞争的解释及能区分它们的观测。提示：若两种解释都预测“平均成功率提高”，增加阶段统计、时间干预或负对照，直到实验能反驳其中之一。

本章要点

研究缺口应落实为可测变量。跨本体成本、覆盖长尾、记忆失效、接触信息、拒绝分母和闭环时序，决定一个结论能否离开原任务与原硬件继续成立。

第二十一章　如何开展一项可信的具身研究

21.1　先区分观测事实与待检验假设

一个研究问题需要可观察现象、候选机制、介入变量、预期结果和反例。例如下面是**尚未运行的假设性方案**，不是本次发现，也不是 OpenVLA 论文结论：

> 假设连续动作模型在某些接触阶段对专家低位宽更敏感。固定观测、数据和控制器，分别量化视觉/语言主干与连续动作专家；再控制延迟，检验误差是否仍集中于接触阶段。若只把全精度模型人为降速就能复现全部退化，数值敏感性解释便受到反驳。

模型结构必须与假设一致：原始 token OpenVLA 无独立连续 action expert，不能在它上面声称做了该模块的消融。OpenVLA 的量化试验 [6] 支持同时检查吞吐和成功率的必要性，不能替这个新假设提供结果。

21.2　让证据范围匹配主张

最小实验先选择一个简单、公开、接口匹配的强基线，再加与新机制最接近的对照。若主张跨任务，留出任务；若主张跨本体，留出本体并报告适配成本；若只研究二维动力学，不必强行包装成通用机器人结论。固定评测初态、成功判据、最大步数、恢复/接管规则和数据划分后再训练。同一个场景的相邻帧不能被当成独立泛化测试样本。

主实验回答是否有收益，组件消融检验哪个机制导致收益，压力测试寻找结论边界。组件替换需保持数据和训练预算可比较；多种变化同时发生时，可做析因设计检查交互，而非把全部增益归给一个组件。探索性选择与最终确认应使用不同数据，否则最好的配置会因反复查看测试集而乐观。

21.3　同时控制数值精度与时间

机器人部署至少有两种公平问题：同一命令时间表下数值变化的影响，以及真实系统各自最高吞吐下的整体收益。它们应分开实验，不能在需要快时固定 Hz、需要高分时又改为各自最大频率。记录

$$\text{outcome} = f(\text{模型精度}, \text{数据}, H_o, H_p, H_e, \text{延迟与队列}, \text{控制器}, \text{初态}).$$

可用全精度人工延迟对照和固定执行时间表诊断时序效应；测量端到端延迟分布与截止时间违反，而非只有 GPU 内核均值。共享初态的比较可按任务/初态配对，多个训练种子的变化与单策略 rollout 的二项不确定性须分别报告。

21.4　用最小实验建立独立参考

随书二维到达算例验证关节与末端接口、单位和时间线；轻量行为克隆消融检验观测/动作历史或 chunk 的作用；合成事件记录演示成功率区间、接管与恢复分类。这些实验验证局部机

制与评测流程，不能替代眞机 VLA 或眞实接触实验。手算、合成教学数据、实际训练结果、论文报告和未运行方案必须分别标记。

代码核验应有独立参考：二维运动学对照手算端点，置信区间对照边界及已知数值，统计分母对照逐条事件记录。重复调用同一个实现得到相同结果只说明可重复，不说明正确。若某个大型仿眞实验因依赖或硬件未运行，应交付冻结配置、命令和成功判据，并明确没有结果；不能用预期数值填表。

21.5 失败记录、版本与阶段性部署

一个可维护研究包保存代码版本、数据 checksum/划分、配置、种子、依赖与硬件、训练预算、checkpoint、评测初态和逐次 rollout 日志。机器人日志再增加相机标定、控制器/固件版本、实际命令、队列、人工干预、失败时间及恢复事件。控制器或硬件变化后的试验应保留分层，不能直接拼成同一性能数字。

从离线字段/坐标核验、固定初态仿眞、压力测试，到影子预测和低速受限眞机，是一种逐步收集证据的路线。每一阶段有自己的退出条件；影子预测没有施加动作，不能证明闭环成功，少量低速试验也不能证明一般物理安全。负结果应保存范围和预算，例如哪种分辨率/冻结策略无收益、是否受数据或吞吐限制，而不写成算法永远无效。

本章要点

可信研究把事实、假设与证据范围分清，冻结协议并保留逐次失败记录。结论应回答机制在什么条件下有效，以及哪一种反证会迫使我们修改解释。

第二十二章　可选研究练习：从机制假设到反证

本章给出六个可缩放的研究练习，不预设读者已有量化、码本或低秩研究积累。完成随书最小实验（第十八章）后，可按资源和兴趣选择一项。每项先验证局部假设，再决定是否投入大型 VLA 与眞机；所有下述结果均待实际实验，不能写成预期增益。

22.1 练习一：混合精度收益来自数值还是时序

假设。某连续动作策略的不同模块对量化误差敏感度不同，保留动作路径精度可能改善固定时序下的闭环表现。动作阶段的重要性可先用离线代理描述：

$$\min_{\mathbf{k}} \ \mathbb{E}_{h_t \sim D_{\text{cal}}} \sum_j w_{t,j} \left\| a^{\text{fp}}_{t+j|t} - a^{\mathbf{k}}_{t+j|t} \right\|^2, \quad M(\mathbf{k}) \le M_0, \quad L(\mathbf{k}) \le L_0.$$

$\mathbf{k}$ 是各模块位宽，D_{cal} 是独立校准集；权重 $w_{t,j}$ 预先定义，M, L 应来自实际显存与延迟测量。这个目标是可计算代理，不是任务失败风险，也不保证代理最小者闭环最好。

最小实验。先在小型连续策略上比较同位宽、动作路径高精度、主干高精度；冻结训练预算、初态与控制器。分别进行固定时间表和各自实际吞吐两组评测，并加全精度人工延迟对照。扩大时可选 OpenVLA-OFT 或 SmolVLA；模型必须确有被研究的动作模块 [14, 15]。

反证与提示。若收益仅由更低延迟解释，不能归因于动作模块数值敏感性；若离线代理排序和闭环排序无关，应放弃该代理或收缩主张。按自由空间、接触、终止分层是诊断，不能事后只挑有利阶段。通用量化内核与部署方法可参阅《智能体模型量化压缩与部署》。

22.2 练习二：动作码本保留了轨迹结构吗

假设。相同码率下，编码整个 chunk 的残差码本比逐维量化保留更多与任务相关的时间结构。可从两级残差 VQ 开始：$A \approx e^{(1)}_{k_1} + e^{(2)}_{k_2}$，同时记录码本利用率、重建误差与动作变化率。不能仅因某个 code 经常对应“接近”就认定已学到可迁移运动 primitive。

最小实验。对同一示范划分比较逐维 bin、普通 VQ、两级残差 VQ；保持有效码率、策略容量与训练预算可比。在本体内留出任务，再做留出本体或接口变换。检查 chunk padding、动作单位和归一化统计，避免码本仅编码数据集 ID。FAST 的频域压缩可作为另一编码思路 [17]，但改变 tokenizer 与改变训练数据须分开。

反证与提示。若重建更好却任务退化，或共享码本仅在见过的本体有效，不能支持物理结构迁移。可在相同初态替换一个 code，检验动作变化是否稳定且可解释；也要记录码本坍缩和不可执行动作，不能只展示成功轨迹。

22.3 练习三：低秩残差能否修正 flow 轨迹

假设。固定量化 flow 策略的主要速度误差位于较低维子空间，小型残差能在有限参数预算内补偿。将 chunk 展平为 $\widetilde{A}^{\tau} \in \mathbb{R}^{d_A}$，$d_A = H_p d_a$，写为

$$v_{\text{corr}}(\widetilde{A}^{\tau}, c, \tau) = v_Q(\widetilde{A}^{\tau}, c, \tau) + B_r U_r \psi(\widetilde{A}^{\tau}, c, \tau),$$

其中 $\psi \in \mathbb{R}^d$，$U_r \in \mathbb{R}^{r\times d}$、$B_r \in \mathbb{R}^{d_A \times r}$，量化模型冻结，只训练残差。输出需恢复为 $H_p \times d_a$；不能把动作矩阵 A 与低秩矩阵重名。

最小实验。在小型 flow head 上比较相同参数量的线性残差、低秩残差与可训练非低秩对照，固定噪声和积分步数。同时测单步速度误差、积分后的 chunk 误差和闭环任务结果，再考虑时间条件化的 rank 或多个时间区间共享参数。

反证与提示。若残差需要接近全维才能有效，或训练集补偿不能迁移到新初态，低秩假设不成立。误差可能来自采样器、动作归一化或时延，应先排除这些接口原因。

22.4 练习四：动态执行长度是否优于固定 chunk

假设。在反馈稀缺的阶段缩短 H_e、在平稳阶段延长 H_e，可在相同计算预算下改善任务结果。先固定预测长度 H_p，只改变执行步数；动态改变分辨率、积分步数和模型规模应留到后续，否则无法定位机制。

最小实验。从第十八章的固定执行长度消融起步，加一个可见状态规则或经验证的风险信号；与固定 H_e、等平均调用数的周期策略和随机切换对照。报告成功率、调用次数、反馈年龄、端到端延迟分位数与截止时间违反。SmolVLA 的异步队列可作为扩大实验的系统入口 [14]。

反证与提示。若随机或周期切换取得相同收益，证据不足以支持信号识别了困难阶段。若动态策略只是调用更多计算，应先匹配预算；预测不确定性未校准时不能把它当作安全概率。

22.5 练习五：恢复能力与失败检测分开学习

假设。对失败状态学习可恢复性及恢复技能，比只训练失败分类更有助于完成整段任务。定义可恢复性为在指定技能集合、剩余预算和约束下恢复成功的概率，而不是状态固有的无条件标签。

最小实验。在可控仿真中生成位置偏移、空抓、遮挡等扰动，分别记录首次失败、检测、恢复尝试与最终结果。比较仅检测后停止、固定重试、条件恢复；同一初态/扰动配对，冻结重试和总时间预算。RoboMIND 的失败标签可作外部监督来源 [39]，使用前核查标签定义与许可，不能将所有失败动作直接当专家。

反证与提示。若恢复成功只来自更长超时，或改善最终成功却增加危险事件，不能支持更可靠部署。必须保留无干预成功、恢复成功、人工辅助与失败的分母，不得删除恢复失败轨迹。

22.6 练习六：压缩预测器是否保留动作排序

假设。对象/接触等任务相关表示可能以更低预算保留候选动作排序，而不需要最高像素保真度。这是表示充分性的假设，不是 mask 或 latent 天然优于 RGB 的结论。

最小实验。在固定候选集上比较未压缩与压缩预测器，记录排序正确率和真实环境中的规划 regret；再检验闭环收益。若把更好的真实模型作为参照，应说明它的成本，不能在部署时偷偷调用参照。世界模型表征、学习与规划细节参阅《世界模型》；本书重点保留动作接口与任务证据。

反证与提示。若视觉指标保留而候选排序变差，压缩目标并未保留决策信息；若排序更好却任务不变，排查候选质量和控制接口。

表 22.1 按已有条件选择起步范围，资源需求不构成创新性评级。

已有条件	可先做的练习	扩展前应获得的证据
CPU 与轻量任务	码本、动态执行、恢复	冻结划分与预算下的局部机制
可训练小型动作头	低秩补偿、压缩预测器	代理指标与闭环指标的关系
GPU 与公开 VLA	精度/时序及动作头消融	实际内核、显存、时延和任务协议
机器人平台与权限	目标本体压力测试	标定、有限执行、接管和逐次事件日志

研究提示

先写一页可被否证的方案：主张、数据划分、固定变量、成功判据、反证条件、预算与停工条件。局部假设不成立也是结果；无需通过追加模块把所有曲线都变成正向。

第二十三章　学习与复现路线：以交付物检查进度

以下用十二周作为可伸缩日程，读者可按前置知识和算力拉长。每阶段以能检查的文件、推导或轨迹结束，而不是以读过多少模型名字结束。基础路线可使用随书轻量实验；大型 VLA、完整仿真基准和真机均是额外分支，没有运行就明确标注。

23.1　第 1–2 周：接口、概率与时间线

先理解状态、观测、历史和 belief，推导有限时域轨迹概率与条件成功概率链。复查关节增量、末端绝对/相对动作、坐标系、单位和控制频率。**交付物：**手算二维到达目标、执行独立运动学参考、画出含采样和推理延迟的时间线，标记 H_o, H_p, H_e。若动作方向或时间起点解释不清，先修复接口，暂不加载大型模型。

23.2　第 3–4 周：建立可复现的行为克隆基线

阅读行为克隆、DAgger 和 ACT/Diffusion Policy 的对应部分 [3, 18, 19]。运行第十八章的轻量消融，检查固定划分、种子、训练预算和成功判据，并重算一次统计表。**交付物：**一条命令、完整配置、实际日志、按任务而非帧划分的数据说明，以及至少一个离线误差和闭环成功不一致的解释。这个阶段不要求复现大型机器人论文的成功率。

23.3　第 5–6 周：动作表示与模型卡

在同一条件数据上实现或阅读回归、token、diffusion/flow 的训练与采样，检查噪声端点、padding mask 和反归一化。按 RT-1、RT-2、OXE、Octo、OpenVLA 的顺序比较接口 [6, 13, 16, 23, 24]，再按需要读 π_0、FAST 或 OFT[15, 17, 22]。**交付物：**两张同字段模型卡、一张预测/执行时间线，以及一个说明动作头适用条件的算例。有 GPU 时可增加 checkpoint 离线推理，记录实际版本和未支持的输入，不把加载成功称为控制成功。

23.4　第 7–8 周：数据与评测契约

将一份示范转换到本书 schema，检查相机/动作时间戳、坐标、夹爪符号、动作统计和 episode 边界。若接入 LIBERO、CALVIN 或 RLBench，先阅读其任务与初态协议，不能把各自成功率放在同一排名中。**交付物：**数据契约、分组划分列表、标定和归一化检查、逐次事件记录与置信区间。大型仿真依赖暂不可用时，保留可执行配置与未运行说明，先完成合成记录和轻量任务验证。

23.5　第 9–10 周：验证、恢复与计算预算

依据第八章实现一个状态谓词验证器，构造成功、遮挡未知、空抓和恢复轨迹。测量传感器到命令生效的路径，比较固定与较短执行时域；若研究量化，增加人工延迟对照。**交付物：**技能

契约、失败分类、无干预/恢复/人工辅助的分开计数，以及延迟分布与预算表。可选阅读 SayCan 和近期可执行技能材料 [4, 38]，重点核查接口与证据，避免照搬特定系统成功率。

23.6 第 11–12 周：一个机制与一份研究包

从第二十二章选择一个假设，先做等预算的最小对照，再决定是否扩大任务或本体。**交付物：**问题陈述、证据定位、冻结配置、组件消融、负结果、失败示例和限制说明。训练种子不足、真机未运行或外部论文尚未复现，都应在结果旁直接标出；不能用预期效果替代测量。

练习与提示

最终自检：方法改变了轨迹分布的哪一部分？它依赖哪些接口、数据与时间假设？什么结果会让你放弃当前解释？提示：如果答案只有“模型更强”，回到变量和负对照，继续收缩问题。

本章要点

学习进度可以由可复现基线、可检查接口和可反驳假设衡量。公式、代码、闭环轨迹与失败记录应彼此对应，资源限制只改变实验规模，不改变证据标准。

第六部分

附录：公式、接口与阅读工具

附录 A　核心公式与使用条件

A.1　行为克隆、chunk 与优势加权

历史 $h_t = (o_{0:t}, a_{0:t-1})$ 包含策略可用信息；有限窗口只保留 H_o 步时，应明确它是近似。单步随机策略与确定性 chunk 回归的常见目标分别为

$$\mathcal{L}_{\mathrm{BC}} = -\mathbb{E}_{(h_t,a_t)\sim D} \log \pi_\theta(a_t \mid h_t),$$
$$\mathcal{L}_{\mathrm{chunk}} = \mathbb{E}_D \frac{\sum_{j=0}^{H_p-1} m_{t,j} \|\hat{a}_{t+j|t} - a_{t+j}\|_1}{\max(1, \sum_j m_{t,j})}.$$

$m_{t,j}$ 表示该未来动作存在且属于同一 episode。训练预测 H_p 步不等于部署执行 H_e 步；padding 不能被当作真实零动作监督。优势加权的一个有截断版本是

$$\mathcal{L}_{\mathrm{AWBC}} = -\mathbb{E}_D\Big[\min\{w_{\max}, \exp(\widehat{A}(h_t,a_t)/T_A)\} \log \pi_\theta(a_t \mid h_t)\Big], \quad T_A > 0.$$

它依赖优势估计与数据覆盖。样本权重不让固定日志包含未访问动作的真实后果，不能消除分布外价值误差。

A.2　SAC：完整状态假设、终止与采样分布

以下采用 Markov 状态 s；部分可观测时应改用有效信息状态或合适的循环估计，不能把单张图像自动当作 Markov 状态。双 critic、非终止 mask $m = 1 - d_{\mathrm{term}}$ 下，一次采样目标可写为

$$a' \sim \pi_\theta(\cdot \mid s'),$$
$$y = r + \gamma m \left[\min_{i=1,2} Q_{\bar{\phi}_i}(s', a') - \alpha \log \pi_\theta(a' \mid s')\right],$$
$$\mathcal{L}_{Q_i} = \mathbb{E}_{(s,a,r,s',d_{\mathrm{term}})\sim D}(Q_{\phi_i}(s,a) - \mathrm{sg}(y))^2,$$
$$\mathcal{L}_\pi = \mathbb{E}_{s\sim D,\, a\sim\pi_\theta(\cdot|s)} \left[\alpha \log \pi_\theta(a \mid s) - \min_i Q_{\phi_i}(s,a)\right].$$

$\bar{\phi}_i$ 为目标网络，$\alpha > 0$ 是熵系数，sg 表示目标不参与 critic 梯度。真正任务终止不 bootstrap；外部采样时间上限不一定是终止。若任务本身有有限时限，应把剩余时间纳入状态并按任务终止处理；纯日志截断只有在下一状态有效等条件下才保留 bootstrap。采用动作 squash 时，密度需计入变量变换；shield 修改后的动作另行记录，不能直接把它当成未变换策略分布的样本。

A.3　Diffusion 与 flow matching 的端点

令 A 为归一化动作 chunk，c 为观测/语言/本体条件。DDPM 式噪声预测的一种约定是

$$A^{(k)} = \sqrt{\bar{\alpha}_k} A + \sqrt{1-\bar{\alpha}_k}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$
$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{A,c,k,\epsilon} \left\| M \odot \left(\epsilon - \epsilon_\theta(A^{(k)}, k, c)\right)\right\|_F^2.$$

M 对有效动作与维度做掩码。采样从高噪声开始沿对应调度去噪；更换 DDIM、噪声参数化或步数后，应同步修改采样公式，不能只替换损失名称。线性条件 flow 的一种约定是

$$A^\tau = (1-\tau)\epsilon + \tau A, \quad 0 \le \tau \le 1,$$
$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left\| M \odot \left(v_\theta(A^\tau, c, \tau) - (A - \epsilon)\right)\right\|_F^2,$$
$$\frac{\mathrm{d}A^\tau}{\mathrm{d}\tau} = v_\theta(A^\tau, c, \tau), \qquad A^0 \sim \mathcal{N}(0, I).$$

在本约定中 $\tau = 0$ 是噪声、$\tau = 1$ 是数据，采样沿正方向积分。采用相反时间定义时速度符号和端点必须同时改变。采样后的动作还要按正确数据源反归一化，并检查单位、坐标、范围与时效。

A.4 MPC 与终端条件

公式 (7.1) 以学到的动态滚动 H_p 步，在可行候选中近似优化，实际只执行 $H_e \le H_p$ 步。预测终止之后不累计后续奖励或终端价值；候选动作约束和状态约束分别作用于真实可表达的量。模型不确定、求解失败或结果过期时应返回明确失败，而不是把空计划视为零动作成功。

A.5 成功率区间与分母

对预先固定协议下 $n > 0$ 次独立 Bernoulli 试验、k 次成功，$\hat{p} = k/n$，Wilson 区间为

$$\frac{\hat{p} + z^2/(2n) \;\pm\; z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}}{1 + z^2/n}.$$

95% 区间通常取 $z \approx 1.96$；$k = 0$ 或 $k = n$ 仍有非零区间宽度，$n = 0$ 不定义。同一训练策略的 rollout 区间不包括训练随机性；共享场景、相关初态或重复测试需要配对或分组分析。无干预成功、恢复成功与人工辅助应各自列出，不能在失败后重置分母。

附录 B　数据、动作与运行接口检查表

本表用于记录是否已有证据。勾选“有控制限制”不构成形式化保证，硬件启用状态和算法假设仍需对应实际测试。

B.1　数据接入前

- ☐ 每个传感器、策略输出和实际动作有时间戳；时钟域、同步误差、插值和丢帧规则已写明。
- ☐ 相机内外参、畸变模型、外参方向与标定日期可追踪；视觉/触觉/本体状态的有效性可区分。
- ☐ q 为关节向量；本体状态 x^{prop} 的位置、速度、力等分量及单位明确。
- ☐ 关节/末端、绝对/增量、工具系/基座系、旋转顺序及夹爪开闭符号有独立样例验证。
- ☐ 采集频率、控制周期、H_o, H_p, H_e 和动作生效延迟分别记录，重采样不改变物理语义。
- ☐ episode ID、终止、超时、截断、padding 和动作有效维 mask 已分开。
- ☐ 先按待测泛化轴冻结场景/任务/对象/操作者/本体分组，再切帧；相邻帧和同次采集不跨集合泄漏。
- ☐ 归一化、分位数、码本和增强参数只从训练集估计，带数据源与版本标识。
- ☐ 示范、失败、接管、恢复和合成片段有 provenance、许可与质量标注，重叠来源可去重。

B.2　闭环执行前

- ☐ 坐标、反归一化和二维参考通过；shape 与单样本拟合只能作辅助检查。
- ☐ 影子预测的方向/幅度与时间起点正确；影子结果尚未被计作闭环成功。
- ☐ 工作空间、速度、力矩、碰撞等控制限制对应实际平台与配置，修改后的命令会被记录。
- ☐ 队列对过期、取消、乱序、断连和超时有明确处置；无计划与零动作有不同结果码。
- ☐ 急停、人工接管和 fallback 已按平台测试；恢复次数与总预算预先冻结。
- ☐ 记录观测年龄、推理/通信/排队延迟、截止时间违反和实际控制频率。
- ☐ 成功/失败/未知的验证器、初态、最大时长和干预计数已冻结，日志可重算统计表。

B.3　结果归档前

- ☐ 代码、环境、种子、划分、训练预算、checkpoint、硬件/控制器/标定版本均可定位。
- ☐ 逐次结果含首次失败、恢复、人工辅助、最终状态与排除原因；排除规则在查看结果前确定。
- ☐ 手算、合成教学、实际训练、论文报告、未运行方案分别标记，没有用预期值填补缺失结果。

附录 C　论文阅读与证据记录模板

每篇论文形成一张可比较的记录卡。缺失字段写“未披露”，不能用另一版本的配置补成确定事实。

1. **身份与版本：**作者、完整题名、年份、DOI/arXiv 及版本、核查日期；保存原始页面或元数据。
2. **问题与边界：**任务、本体、环境、观察输入、动作语义、坐标/单位、H_o, H_p, H_e。
3. **机制与假设：**模型表示、动作头、损失与采样；作者认为瓶颈是什么，依赖哪些可观测性或数据条件。
4. **数据：**来源、本体、场景/任务覆盖、数量单位、成功/失败、许可、混合权重和训练测试划分。
5. **训练：**初始化、阶段、冻结模块、优化器、预算、种子和消融；规模估算与实测耗时分开。
6. **部署：**硬件、精度、控制器、延迟分布、队列、实际频率、超时/接管/恢复策略。
7. **证据定位：**每个机制或数值对应页、节、表、图或代码 commit；论文自报与独立复现分别列出。
8. **评测协议：**初态、任务数、rollout 数、成功判据、统计单位、区间和预算；基准是否相同。
9. **公开资产：**代码、checkpoint、config、数据版本及可运行范围；有项目网页不等于完整可复现。
10. **限制与下一问：**哪些轴没测，哪种反例会推翻解释，最小额外对照是什么。

遇到“显著提高”“实时”“通用”等词，先转换成可核查的问题：相对哪个基线、什么预算/统计检验、哪段延迟、哪些留出任务与本体？跨协议表格可以描述设计，不能自动构成性能排名。引用支持不到当前主张时，缩小或删除主张；仅增加参考文献数不能修复证据缺口。

附录 D　术语与符号表

术语	本书用法
Embodiment	机器人的身体、传感器、执行器、控制接口及其物理约束。
VLA	Vision-Language-Action，以视觉/语言等条件产生机器人动作；具体观测通道随模型而异。
机器人基础模型	多任务、场景或本体数据上预训练并适配下游的模型；“基础”不表示普适成功保证。
Action chunk	一次预测的未来动作序列，长度 H_p；实际执行长度 H_e 可不同。
Temporal ensemble	融合重叠 chunk 对同一执行时刻的预测，需要正确时间对齐。
Flow matching	学习连接噪声与数据分布的速度场；时间方向由具体约定确定。
Diffusion Policy	通过迭代去噪生成条件动作序列的策略。
World model	预测动作作用下状态、观测或任务相关表示演化的模型。
Affordance	对技能在当前条件下可执行性或成功可能性的描述；须说明二值规则、价值或概率口径。
Sim2Real	将仿真经验迁至真实机器人，涉及动力学、观测、动作与控制系统差异。
跨本体	在不同身体、动作空间或控制器之间复用结构/经验；目标接口适配仍需报告。
Offline RL	仅从固定日志学习价值或策略，不继续在线采样；受覆盖与分布外估计限制。
Post-training	预训练后用目标示范、奖励或偏好改变行为，监督学习与 RL 机制不可混同。
Runtime shield	按给定模型/规则拦截、投影或替换动作；保证范围依赖建模、可行性、延迟和执行器假设。
控制约束	控制器实施的幅度、速度、力或状态限制；启用限制不等于完成形式化证明。
形式化保证	在明确系统模型、假设与性质下的数学保证，不等同于有限压力测试通过。

术语	本书用法
Embodiment gap	从可复用模型/表示/数据到目标机器人执行之间尚需完成的适配工作。
接管与恢复	接管涉及人工介入；恢复可以自主进行，二者均计入整段轨迹与预算。

表 D.2 核心符号，局部算法的额外符号在正文定义。

符号	含义
s_t, o_t, a_t	真实状态、观测与动作，不能默认三者维度或可观测性相同
h_t	到时刻 t 的观测与历史动作；有限窗口是其近似
b_t	$p(s_t \mid h_t)$，信念状态；学习表示 z_t 不自动等于它
q_t, x_t^{prop}	关节向量；包含关节、速度、力等的本体状态向量
H_o, H_p, H_e	观测历史步数、预测步数、每次实际执行步数
$a_{t+j\|t}, A_t$	在 t 时刻预测的第 j 个动作及其 chunk
d_{term}, m	真终止标记与 bootstrap mask $m = 1 - d_{\text{term}}$
γ, α	折扣与 SAC 熵系数；diffusion 调度系数在对应公式另作区分

结语

机器人基础模型把语义先验带入行动，但一个完整结果仍要经过坐标、时钟、控制器、数据和评测协议的检验。读懂模型以后，下一步是把一个假设变成可复现的小实验：先确定接口，再冻结数据与判据，保存失败，最后用闭环证据决定是否继续。

参考文献

[1] Y. Domae et al. *The Embodiment Gap in Robot Foundation Models*. Version v1. 2026. arXiv: 2608.18433. URL: https://arxiv.org/abs/2608.18433v1 (visited on 10/22/2026).

[2] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. "Planning and acting in partially observable stochastic domains". In: *Artificial Intelligence* 101.1-2 (1998), pp. 99–134. DOI: 10.1016/S0004-3702(98)00023-X. (Visited on 10/22/2026).

[3] S. Ross, G. J. Gordon, and J. A. Bagnell. *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning*. Version v3. 2010. arXiv: 1011.0686. URL: https://arxiv.org/abs/1011.0686v3 (visited on 10/22/2026).

[4] M. Ahn et al. *Do As I Can, Not As I Say: Grounding Language in Robotic Affordances*. Version v2. 2022. arXiv: 2204.01691. URL: https://arxiv.org/abs/2204.01691v2 (visited on 10/22/2026).

[5] NVIDIA et al. *GR00T N1: An Open Foundation Model for Generalist Humanoid Robots*. Version v2. 2025. arXiv: 2503.14734. URL: https://arxiv.org/abs/2503.14734v2 (visited on 10/22/2026).

[6] M. J. Kim et al. *OpenVLA: An Open-Source Vision-Language-Action Model*. Version v3. 2024. arXiv: 2406.09246. URL: https://arxiv.org/abs/2406.09246v3 (visited on 10/22/2026).

[7] M. Shridhar, L. Manuelli, and D. Fox. *Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation*. Version v2. 2022. arXiv: 2209.05451. URL: https://arxiv.org/abs/2209.05451v2 (visited on 10/22/2026).

[8] A. Zeng et al. *Transporter Networks: Rearranging the Visual World for Robotic Manipulation*. Version v3. 2020. arXiv: 2010.14406. URL: https://arxiv.org/abs/2010.14406v3 (visited on 10/22/2026).

[9] M. Shridhar, L. Manuelli, and D. Fox. *CLIPort: What and Where Pathways for Robotic Manipulation*. Version v1. 2021. arXiv: 2109.12098. URL: https://arxiv.org/abs/2109.12098v1 (visited on 10/22/2026).

[10] Y. Jiang et al. *VIMA: General Robot Manipulation with Multimodal Prompts*. Version v2. 2022. arXiv: 2210.03094. URL: https://arxiv.org/abs/2210.03094v2 (visited on 10/22/2026).

[11] D. Driess et al. *PaLM-E: An Embodied Multimodal Language Model*. Version v1. 2023. arXiv: 2303.03378. URL: https://arxiv.org/abs/2303.03378v1 (visited on 10/22/2026).

[12] H.-S. Fang et al. *RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot*. Version v2. 2023. arXiv: 2307.00595. URL: https://arxiv.org/abs/2307.00595v2 (visited on 10/22/2026).

[13] A. Brohan et al. *RT-1: Robotics Transformer for Real-World Control at Scale*. Version v2. 2022. arXiv: 2212.06817. URL: https://arxiv.org/abs/2212.06817v2 (visited on 10/22/2026).

[14] M. Shukor et al. *SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics*. Version v1. 2025. arXiv: 2506.01844. URL: https://arxiv.org/abs/2506.01844v1 (visited on 10/22/2026).

[15] M. J. Kim, C. Finn, and P. Liang. *Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success*. Version v2. 2025. arXiv: 2502.19645. URL: https://arxiv.org/abs/2502.19645v2 (visited on 10/22/2026).

[16] A. Brohan et al. *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control*. Version v1. 2023. arXiv: 2307.15818. URL: https://arxiv.org/abs/2307.15818v1 (visited on 10/22/2026).

[17] K. Pertsch et al. *FAST: Efficient Action Tokenization for Vision-Language-Action Models*. Version v1. 2025. arXiv: 2501.09747. URL: https://arxiv.org/abs/2501.09747v1 (visited on 10/22/2026).

[18] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. *Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware*. Version v1. 2023. arXiv: 2304.13705. URL: https://arxiv.org/abs/2304.13705v1 (visited on 10/22/2026).

[19] C. Chi et al. *Diffusion Policy: Visuomotor Policy Learning via Action Diffusion*. Version v5. 2023. arXiv: 2303.04137. URL: https://arxiv.org/abs/2303.04137v5 (visited on 10/22/2026).

[20] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. Version v2. 2020. arXiv: 2006.11239. URL: https://arxiv.org/abs/2006.11239v2 (visited on 10/22/2026).

[21] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. *Flow Matching for Generative Modeling*. Version v2. 2022. arXiv: 2210.02747. URL: https://arxiv.org/abs/2210.02747v2 (visited on 10/22/2026).

[22] K. Black et al. π_0*: A Vision-Language-Action Flow Model for General Robot Control*. Version v4. 2024. arXiv: 2410.24164. URL: https://arxiv.org/abs/2410.24164v4 (visited on 10/22/2026).

[23] Open X-Embodiment Collaboration et al. *Open X-Embodiment: Robotic Learning Datasets and RT-X Models*. Version v9. 2023. arXiv: 2310.08864. URL: https://arxiv.org/abs/2310.08864v9 (visited on 10/22/2026).

[24] Octo Model Team et al. *Octo: An Open-Source Generalist Robot Policy*. Version v2. 2024. arXiv: 2405.12213. URL: https://arxiv.org/abs/2405.12213v2 (visited on 10/22/2026).

[25] Physical Intelligence et al. $\pi_{0.5}$*: a Vision-Language-Action Model with Open-World Generalization*. Version v1. 2025. arXiv: 2504.16054. URL: https://arxiv.org/abs/2504.16054v1 (visited on 10/22/2026).

[26] Physical Intelligence et al. $\pi_{0.7}$*: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities*. Version v2. 2026. arXiv: 2604.15483. URL: https://arxiv.org/abs/2604.15483v2 (visited on 10/22/2026).

[27] S. Liu et al. *RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation*. Version v2. 2024. arXiv: 2410.07864. URL: https://arxiv.org/abs/2410.07864v2 (visited on 10/22/2026).

[28] Q. Li et al. *CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation*. Version v1. 2024. arXiv: 2411.19650. URL: https://arxiv.org/abs/2411.19650v1 (visited on 10/22/2026).

[29] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng. *DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control*. Version v3. 2025. arXiv: 2502.05855. URL: https://arxiv.org/abs/2502.05855v3 (visited on 10/22/2026).

[30] Gemini Robotics Team et al. *Gemini Robotics: Bringing AI into the Physical World*. Version v1. 2025. arXiv: 2503.20020. URL: https://arxiv.org/abs/2503.20020v1 (visited on 10/22/2026).

[31] D. Ha and J. Schmidhuber. *World Models*. Version v4. 2018. DOI: 10.5281/zenodo.1207631. arXiv: 1803.10122. URL: https://arxiv.org/abs/1803.10122v4 (visited on 10/22/2026).

[32] J. Schrittwieser et al. *Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model*. Version v2. 2019. DOI: 10.1038/s41586-020-03051-4. arXiv: 1911.08265. URL: https://arxiv.org/abs/1911.08265v2 (visited on 10/22/2026).

[33] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. *Mastering Diverse Domains through World Models*. Version v2. 2023. arXiv: 2301.04104. URL: https://arxiv.org/abs/2301.04104v2 (visited on 10/22/2026).

[34] N. Hansen, H. Su, and X. Wang. *TD-MPC2: Scalable, Robust World Models for Continuous Control*. Version v2. 2023. arXiv: 2310.16828. URL: https://arxiv.org/abs/2310.16828v2 (visited on 10/22/2026).

[35] J. Bruce et al. *Genie: Generative Interactive Environments*. Version v1. 2024. arXiv: 2402.15391. URL: https://arxiv.org/abs/2402.15391v1 (visited on 10/22/2026).

[36] S. L. Li et al. *Turning Video Models into Generalist Robot Policies*. Version v1. 2026. arXiv: 2605.27817. URL: https://arxiv.org/abs/2605.27817v1 (visited on 10/22/2026).

[37] Y. Lou et al. *Mask World Model: Predicting What Matters for Robust Robot Policy Learning*. Version v2. 2026. arXiv: 2604.19683. URL: https://arxiv.org/abs/2604.19683v2 (visited on 10/22/2026).

[38] W. Wang et al. *EmbodiedSkills: A Unified Framework for Orchestrating, Training, and Deploying VLA Agents*. Version v1. 2026. arXiv: 2609.01281. URL: https://arxiv.org/abs/2609.01281v1 (visited on 10/22/2026).

[39] K. Wu et al. *RoboMIND: Benchmark on Multi-embodiment Intelligence Normative Data for Robot Manipulation*. Version v3. 2024. DOI: 10.15607/RSS.2025.XXI.152. arXiv: 2412.13877. URL: https://arxiv.org/abs/2412.13877v3 (visited on 10/22/2026).

[40] J. Ho and S. Ermon. *Generative Adversarial Imitation Learning*. Version v1. 2016. arXiv: 1606.03476. URL: https://arxiv.org/abs/1606.03476v1 (visited on 10/22/2026).

[41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*. Version v2. 2018. arXiv: 1801.01290. URL: https://arxiv.org/abs/1801.01290v2 (visited on 10/22/2026).

[42] D. Kalashnikov et al. *QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation*. Version v3. 2018. arXiv: 1806.10293. URL: https://arxiv.org/abs/1806.10293v3 (visited on 10/22/2026).

[43] I. Kostrikov, A. Nair, and S. Levine. *Offline Reinforcement Learning with Implicit Q-Learning*. Version v1. 2021. arXiv: 2110.06169. URL: https://arxiv.org/abs/2110.06169v1 (visited on 10/22/2026).

[44] Y. Wang et al. *Learning While Deploying: Fleet-Scale Reinforcement Learning for Generalist Robot Policies*. Version v3. 2026. arXiv: 2605.00416. URL: https://arxiv.org/abs/2605.00416v3 (visited on 10/22/2026).

[45] L. Cao, R. Chen, L. Li, P. Wang, M. Peng, and Y. Li. *Z-1: Efficient Reinforcement Learning for Vision-Language-Action Models*. Version v1. 2026. arXiv: 2606.31846. URL: https://arxiv.org/abs/2606.31846v1 (visited on 10/22/2026).

[46] Y. Qiu, H. Ye, S. Sun, Z. Huang, R. X. Xu, and M. Sun. *GRAFT: Grounded and Efficient Online Reinforcement Adaptation for Fine-Grained Robot Manipulation*. Version v2. 2026. arXiv: 2608.27079. URL: https://arxiv.org/abs/2608.27079v2 (visited on 10/22/2026).

[47] F. Lin et al. *A Systematic Study of Data Modalities and Strategies for Co-training Large Behavior Models for Robot Manipulation*. Version v1. 2026. arXiv: 2602.01067. URL: https://arxiv.org/abs/2602.01067v1 (visited on 10/22/2026).

[48] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World*. Version v1. 2017. arXiv: 1703.06907. URL: https://arxiv.org/abs/1703.06907v1 (visited on 10/22/2026).

[49] S. Nasiriany et al. *RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots*. Version v1. 2024. arXiv: 2406.02523. URL: https://arxiv.org/abs/2406.02523v1 (visited on 10/22/2026).

[50] A. Mandlekar et al. *MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations*. Version v1. 2023. arXiv: 2310.17596. URL: https://arxiv.org/abs/2310.17596v1 (visited on 10/22/2026).

[51] Y. Ye et al. *Data Pyramid for Embodied Manipulation: A Survey*. Version v2. 2026. arXiv: 2607.24744. URL: https://arxiv.org/abs/2607.24744v2 (visited on 10/22/2026).

[52] S. Dasari et al. *RoboNet: Large-Scale Multi-Robot Learning*. Version v2. 2019. arXiv: 1910.11215. URL: https://arxiv.org/abs/1910.11215v2 (visited on 10/22/2026).

[53] H. Walke et al. *BridgeData V2: A Dataset for Robot Learning at Scale*. Version v3. 2023. arXiv: 2308.12952. URL: https://arxiv.org/abs/2308.12952v3 (visited on 10/22/2026).

[54] A. Khazatsky et al. *DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset*. Version v2. 2024. arXiv: 2403.12945. URL: https://arxiv.org/abs/2403.12945v2 (visited on 10/22/2026).

[55] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard. *CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks*. Version v4. 2021. arXiv: 2112.03227. URL: https://arxiv.org/abs/2112.03227v4 (visited on 10/22/2026).

[56] B. Liu et al. *LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning*. Version v2. 2023. arXiv: 2306.03310. URL: https://arxiv.org/abs/2306.03310v2 (visited on 10/22/2026).

[57] M. Savva et al. *Habitat: A Platform for Embodied AI Research*. Version v2. 2019. arXiv: 1904.01201. URL: https://arxiv.org/abs/1904.01201v2 (visited on 10/22/2026).

[58] A. Szot et al. *Habitat 2.0: Training Home Assistants to Rearrange their Habitat*. Version v2. 2021. arXiv: 2106.14405. URL: https://arxiv.org/abs/2106.14405v2 (visited on 10/22/2026).

[59] M. Deitke et al. *ProcTHOR: Large-Scale Embodied AI Using Procedural Generation*. Version v1. 2022. arXiv: 2206.06994. URL: https://arxiv.org/abs/2206.06994v1 (visited on 10/22/2026).

[60] M. Shridhar et al. *ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks*. Version v2. 2019. arXiv: 1912.01734. URL: https://arxiv.org/abs/1912.01734v2 (visited on 10/22/2026).

[61] C. Li et al. *BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation*. Version v1. 2024. arXiv: 2403.09227. URL: https://arxiv.org/abs/2403.09227v1 (visited on 10/22/2026).

[62] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison. *RLBench: The Robot Learning Benchmark & Learning Environment*. Version v1. 2019. arXiv: 1909.12271. URL: https://arxiv.org/abs/1909.12271v1 (visited on 10/22/2026).

[63] J. Gu et al. *ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills*. Version v1. 2023. arXiv: 2302.04659. URL: https://arxiv.org/abs/2302.04659v1 (visited on 10/22/2026).

[64] E. B. Wilson. "Probable Inference, the Law of Succession, and Statistical Inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212. DOI: 10.1080/01621459.1927.10502953. (Visited on 10/22/2026).

[65] Q. Li et al. *Vision-Language-Action Safety: Threats, Challenges, Evaluations, and Mechanisms*. Version v2. 2026. arXiv: 2604.23775. URL: https://arxiv.org/abs/2604.23775v2 (visited on 10/22/2026).

[66] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. *Control Barrier Function Based Quadratic Programs for Safety Critical Systems*. Version v2. 2016. DOI: 10.1109/TAC.2016.2638961. arXiv: 1609.06408. URL: https://arxiv.org/abs/1609.06408v2 (visited on 10/22/2026).