ACT 与 ALOHA：低成本双臂精细操作的完整链路

Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware

Tony Z. Zhao, Vikash Kumar, Sergey Levine, Chelsea Finn

Robotics: Science and Systems 2023
底本 arXiv:2304.13705v1 · 2023-04-23 · 原文 18 页

> **中文深读版 / Original study companion**
> 本册为原创方法与实验解读，配套中英对照 Markdown；不是原文逐段译本。原文采用 arXiv 非独占分发许可，未提供公开翻译授权。英语释义、数学重述、数据辨析与读者推演分开标识，原文全文见同目录官方 PDF。

阅读主线：从观测与动作的接口出发，沿训练目标、生成计算、闭环执行和实验协议逐层追踪，最后连接流匹配与事件观测。

来源页码均指随附英文 PDF。数据为论文报告值；“读者推演”是面向复现的分析建议。

章节导航

1. 1. 问题入口：低成本精度如何转化为任务精度
2. 2. ALOHA 的结构：关节空间映射而非手部重定向
3. 3. 动作监督的语义：目标关节角不能偷换成实测角
4. 4. 动作分块：降低决策频次而非消灭误差
5. 5. 时间集成：融合针对同一时刻的多个预测
6. 6. 条件 VAE：训练时解释变化，测试时固定风格
7. 7. 模型张量逐层追踪
8. 8. 八项任务：感知、接触与多阶段误差
9. 9. 数据与比较协议：十分钟意味着什么
10. 10. 定量结果与消融：不要把百分点说成百分比
11. 11. 附录与局限：失败案例同样是系统规格
12. 12. 连接扩散策略、流匹配与事件观测

配套文件

paper.md：原创中英双语对照
source_map.json：逐块原文位置与覆盖边界
translation_notes.md：符号、术语、原稿差异与课题联动
assets/：原图局部高清裁图
work/：内容源、静态公式、HTML、打印脚本与检查证据

1. 问题入口：低成本精度如何转化为任务精度

S001 · 英文 PDF p.1, 2

论文把两件事作为一个系统来做：ALOHA 解决低成本、高质量双臂演示的采集，ACT 解决如何从这些演示学到细致的视觉闭环。透明杯盖、遥控器电池、魔术贴环都要求毫米级修正，而低成本机械臂的绝对定位误差可达 5–8 mm。系统没有先建立高精度几何世界模型，而是用视觉反馈学习如何补偿误差。

S002 · 英文 PDF p.1, 2

摘要中的 80–90% 是对代表性任务的概括，不是六项真实任务都达到这一水平。表格中的穿魔术贴最终成功率只有 20%，准备胶带为 64%，而电池插入为 96%。同样，“十分钟演示”指记录的有效演示时长；加上复位与操作失误，数据采集占用的墙钟时间是 30–60 分钟。

2. ALOHA 的结构：关节空间映射而非手部重定向

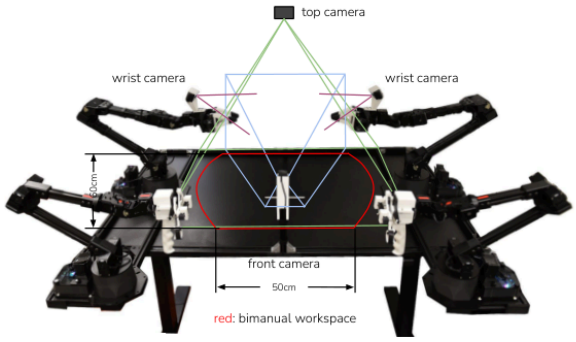
S003 · 英文 PDF p.3, 4

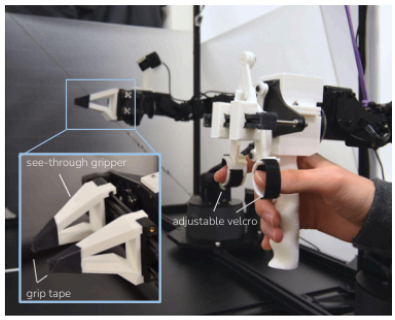
ALOHA 用两台较小的 WidowX 作 leader，两台 ViperX 作 follower。操作者反驱 leader，关节角直接映射给 follower，因此避开了把任意手势转换为末端位姿再做逆运动学的步骤。相似运动学结构在奇异位形附近尤其有价值；但它也把可达空间、关节限制与机构形态绑定在一起，并非对任意异构机械臂都成立。

S004 · 英文 PDF p.3, 4

论文的五项设计原则是低成本、通用、易用、可修复、易搭建。可更换 Dynamixel 电机、3D 打印的透明夹指、增摩胶带、“手柄与剪刀”式反驱机构和橡皮筋重力补偿都服务于持续采集，而不只是硬件展示。四台 RGB 摄像头提供顶部、正面、左右腕部视角。遥操作与数据记录为 50 Hz，附录摄像头实际流为 30 fps，两种采样率应分别记录。

硬件量	原文值	解释
follower 自由度	6 + 夹爪 / 臂	双臂动作 14 维
臂展 / 工作负载	1500 mm / 750 g	额定规格不是任意姿态的保证
重复定位 / 绝对精度	1 mm / 5–8 mm	两个指标不可互换
总预算	低于 20,000 USD	论文 2023 年配置，不是当前报价
RGB / 控制	480×640，30 fps / 50 Hz	不同数据流需时间对齐





ViperX 6dof Arm (follower)

#Dofs	6+gripper
Reach	750mm
Span	1500mm
Repeatability	1mm
Accuracy	5-8mm
Working Payload	750g

原图 3：相机、反驱机构与 follower 规格 · 英文 PDF p.3
左图展示互补视角，中图为操作接口与夹指，右图为机械臂规格。重复定位精度与绝对精度是不同量。

3. 动作监督的语义：目标关节角不能偷换成实测角

S005 · 英文 PDF p.4

训练动作取 leader 关节角，观测中的关节角取 follower 实际位置。这一分工很重要：目标与实测之间的差值通过底层 PID 影响施力，接触物体时 follower 可能尚未到达目标。如果把实测角当作动作标签，网络学习到的就可能是已经被接触约束截断的运动，而不是人类施加的意图。

S006 · 英文 PDF p.4, 6

每步动作是两臂下一时刻的绝对目标角，维度为 14；一个长度 k 的动作块就是 $k \times 14$。底层电机控制器负责高频跟踪。ACT 不直接输出电机力矩，也没有把接触力显式作为监督变量。复现数据格式首先应把“状态”“命令”“生效时间”和单位分开。

$$o_t = (I_t^{(1)}, I_t^{(2)}, I_t^{(3)}, I_t^{(4)}, q_t^{\text{follower}}), \qquad A_t \in \mathbb{R}^{k \times 14}$$

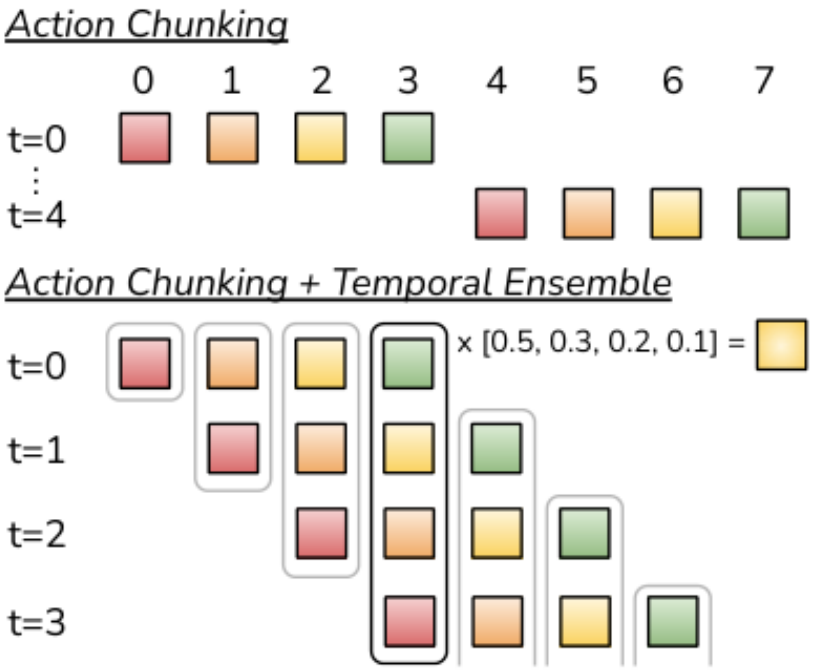
4. 动作分块：降低决策频次而非消灭误差

S007 · 英文 PDF p.4, 5, 9

最简单的分块策略每隔 k 步取一次观测，生成并执行接下来的 k 个动作。它把长轨迹的高层决策次数从约 T 次变为约 T/k 次，还能把“短暂停顿后继续运动”放进一个输出块。这个解释针对有效决策时域；它不是关于误差必然缩小 k 倍的定理，块内运动仍受模型偏差、低层跟踪与外部扰动影响。

S008 · 英文 PDF p.9, 10

在关闭时间集成的消融中，四种模拟设置平均成功率由 $k = 1$ 的约 1% 上升到 $k = 100$ 的约 44%，更大的 200、400 又下降。改善不只来自 Transformer：给 BC-ConvMLP 扩大输出维度、给 VINN 检索动作序列，也能获得收益。过长动作块削弱反应性，并增加长序列拟合难度。



原图 5：动作分块与同一时刻的时间集成 · 英文 PDF p.4
每一行表示一次预测，纵向聚合才是对同一执行时刻的多个估计。不要把该操作改成相邻执行时刻的平均。

5. 时间集成：融合针对同一时刻的多个预测

S009 · 英文 PDF p.5

正式推理可在每个控制步重新查询策略。时刻 t 的动作由多个过去生成块中、都指向 t 的条目融合而成。这不是对相邻时刻的实际执行动作做滑动平均。相同时间目标的融合能削弱预测噪声，同时更平滑地引入新观测，代价是每步都要运行视觉与策略。

S010 · 英文 PDF p.5

原文规定索引 $i = 0$ 是最旧预测，权重为 e^{-mi}。因此 $m > 0$ 时最旧项权重最大，后续较新预测权重逐渐变小；较小的 m 让新预测相对更快进入融合。常见的“新预测权重更大”直觉不能直接套到这个索引定义。实现若反转 FIFO 顺序，公式的效果也会反转，必须明确排序与权重的配对。

S011 · 英文 PDF p.9, 10

时间集成对 ACT 的增益约为 3.3 个百分点、对 BC-ConvMLP 为 4 个百分点，但 VINN 出现下降。原文解释为参数模型受益于消除拟合噪声，而 VINN 检索的是真实演示动作。融合并非总能改进：若不同块代表互斥接触策略，其均值也可能无效，仍需监测动作约束与阶段切换。

$$\hat{a}_t = \frac{\sum_{i=0}^{n_t-1} e^{-mi}\, \hat{a}_t^{(i)}}{\sum_{i=0}^{n_t-1} e^{-mi}}, \qquad i = 0 \text{ denotes the oldest prediction}$$

6. 条件 VAE：训练时解释变化，测试时固定风格

S012 · 英文 PDF p.5, 6, 9, 10

CVAE 编码器看当前本体感觉与未来动作块，输出风格变量 z 的对角高斯分布；它不看图像，以降低训练成本。解码器才接收四视角图像、当前关节位置和 z。训练时，潜变量可解释同一观测下的人类动作差异；测试时编码器被丢弃，并固定 $z = 0$，所以给定同一输入的默认推理是确定性的。

S013 · 英文 PDF p.5, 6

下式采用正文与实现说明中的 L1 重建损失，并将 KL 项展开为对角高斯的闭式形式。β 控制潜变量携带的信息量，而不是直接控制动作平滑。原文算法 1 第 9 行却写 MSE，与正文第 6 页的 L1 存在差异；这里明确保留这个疑点，实际复现应核对代码配置。

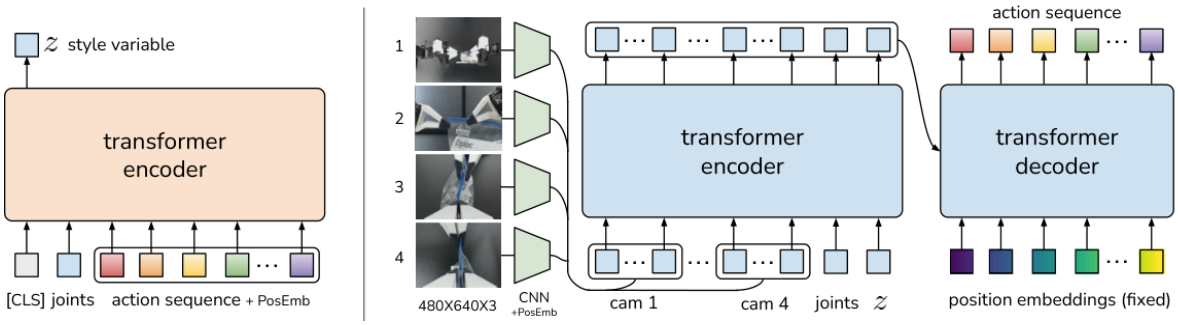
S014 · 英文 PDF p.9, 10

去掉 CVAE 对脚本数据影响很小，却明显损害人类数据表现，说明处理演示中的变化确有价值。但是固定 $z = 0$ 不等于计算动作分布的严格均值，也不能保证覆盖所有任务模式。将 ACT 视为“可随机多样采样的测试策略”时，必须说明是否更改了默认潜变量设置。

$$z = \mu_\phi + \sigma_\phi \odot \eta, \quad \eta \sim \mathcal{N}(0, I)$$

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|A_t,q_t)}[\|A_t - f_\theta(o_t, z)\|_1] + \beta\, D_{\mathrm{KL}}(q_\phi(z \mid A_t, q_t)\|\mathcal{N}(0, I))$$

$$D_{\mathrm{KL}} = \tfrac{1}{2}\sum_j \left(\mu_j^2 + \sigma_j^2 - 1 - \log \sigma_j^2\right)$$



原图 4：ACT 训练编码器与策略解码器 · 英文 PDF p.4
左侧 CVAE 编码器仅训练时存在；右侧才是测试时的策略。潜变量为零是默认推理约定，不是无条件随机采样。

7. 模型张量逐层追踪

S015 · 英文 PDF p.5, 6, 14, 17, 18

每张 480×640 图像经 ResNet-18 得到空间特征，正文按 $15 \times 20 \times 512$ 描述；展平后每视角 300 个 token，四视角合计 1200 个。再加入当前关节与潜变量两个 token，Transformer 编码器输入为 1202×512。动作查询是长度 k 的固定位置嵌入，通过交叉注意力读取编码特征，再投影成 $k \times 14$ 的目标角。

S016 · 英文 PDF p.6, 17, 18

结构细节也有原稿疑点：附录图 11 的部分视觉特征标为 728 维，但正文写 ResNet-18 的 512 维，且图中线性层再投影到 512。应把它记录为图文不一致，不据此杜撰一个 728 通道的标准 ResNet-18。正文给出约 80M 参数、单张 11 GB RTX 2080 Ti 训练约 5 小时、推理约 0.01 s；这是特定硬件和配置的报告值，不包括完整部署时延证明。

参数	论文表 III
学习率 / batch size	1e-5 / 8
编码器层 / 解码器层	4 / 7
隐藏维 / 前馈维 / 头数	512 / 3200 / 8
动作块长度 k	100
潜变量维度	32（图 11）
KL 权重 / dropout	10 / 0.1

8. 八项任务：感知、接触与多阶段误差

S017 · 英文 PDF p.6, 7, 8

真实任务包含拉开立起的密封袋、插电池、开杯盖、穿魔术贴、准备胶带和穿鞋。它们并不是同一种抓放模板：有的需要一臂固定物体、另一臂施力；有的需要空中交接；有的需要在变形后重新定位。透明材料、黑色背景和小目标让 RGB 定位本身成为瓶颈，增加网络容量并不会直接修复缺失的可观测信息。

S018 · 英文 PDF p.6, 8

两个 MuJoCo 任务分别是双臂传方块与空中插接。前者交接间隙约 1 cm，后者插入间隙约 5 mm；真实穿魔术贴的环口为 3 mm×25 mm，带厚约 2 mm。多个阶段相乘会放大前期小误差，因此论文同时报告分阶段成功率，而不只报告最终是否完成。

真实任务	完整成功率	最终判据
Slide Ziploc	88%	完成拉开
Slot Battery	96%	电池完全插入
Open Cup	84%	杯盖打开
Thread Velcro	20%	带端穿过小环
Prep Tape	64%	胶带挂到箱边
Put On Shoe	92%	双夹爪释放后鞋仍固定

9. 数据与比较协议：十分钟意味着什么

S019 · 英文 PDF p.8, 9, 15

每个真实任务通常采 50 条演示，穿魔术贴为 100 条；单条有效操作 8–14 s，对应 50 Hz 下的 400–700 个控制步。两项模拟任务分别采脚本与人类数据，每类 50 条成功演示。脚本控制几乎没有人类停顿与空中交接位置变化，因此在脚本数据上调好的单步模型不应直接被当作人类数据的充分基线。

S020 · 英文 PDF p.8, 9, 15

模拟结果跨 3 个种子、每种子 50 次评估；真实结果为 1 个种子、25 次评估。因此真实成功率最小变化单位是 4 个百分点，且不能从单种子结果分离训练随机性。基线包括 BC-ConvMLP、BeT、RT-1 和 VINN；论文调整了历史长度、隐藏维和检索状态权重，但模型的预训练、感知耦合和动作离散化仍存在多重差异。

S021 · 英文 PDF p.8, 9

真实任务的分阶段数值是累计完成率，不是条件成功概率。例如穿魔术贴 92%→40%→20%，意味着从总体评估到达各阶段的比例；若要估算最后插接阶段的条件成功率，应在相同试验集合下计算 $20/40 = 50\%$，而不能把 20% 直接当作“已抓稳后插入的概率”。

$$\hat{p}(\text{insert} \mid \text{grasp}) = \frac{\#\text{complete insertion}}{\#\text{complete grasp}} = \frac{20}{40} = 0.5$$

10. 定量结果与消融：不要把百分点说成百分比

S022 · 英文 PDF p.8, 9, 10

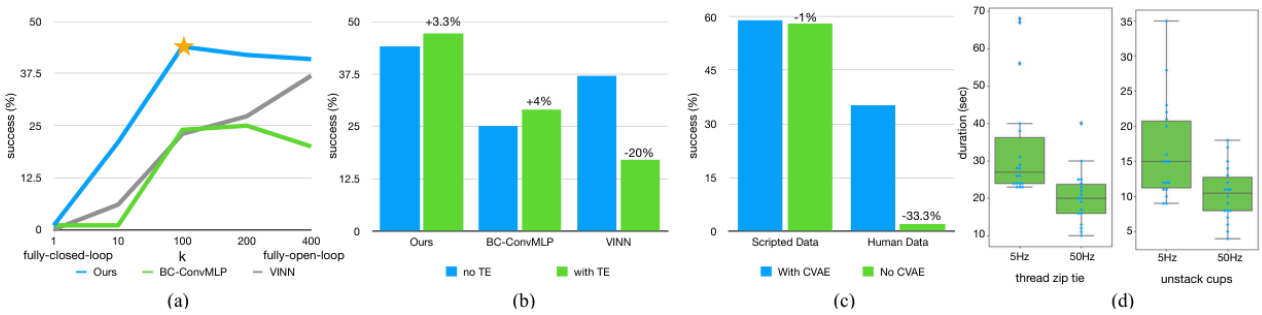
在模拟传方块中，ACT 对脚本/人类数据的最终成功率为 86%/50%；空中插接为 32%/20%。最强先前方法相应最高约为 27%/1% 与 3%/0%。因此正文所说的 59%、49%、29%、20% 差距按表格应理解为百分点差，不能改算成同样数值的相对提升。

S023 · 英文 PDF p.9, 10

消融支持分块与 CVAE 对人类演示的重要性，却也显示组件收益依赖模型类型。不能从“去掉 CVAE 下降”推出“任何噪声都应由潜变量吸收”，因为错标、时序错位和传感异常属于数据质量问题。有效实验应先修复输入与监督对齐，再研究潜变量能否建模合法的人类风格差异。

S024 · 英文 PDF p.10, 15

六位参与者的遥操作研究对比 5 Hz 与 50 Hz，任务为穿扎带和分离叠杯。5 Hz 版本仍向 follower 发送 50 Hz 插值命令，因此对比的是操作者输入更新速度，而不是电机是否以 5 Hz 伺服。研究展示高频闭环对细操作的帮助，但样本规模和任务范围有限。



原图 8：分块、时间集成、CVAE 与更新率消融 · 英文 PDF p.10

各面板回答不同问题。先读横轴与数据来源，再比较数值；时间集成会损害 VINN，组件收益不是普遍单调的。

11. 附录与局限：失败案例同样是系统规格

S025 · 英文 PDF p.14, 15, 16

ALOHA 的平行夹爪降低了成本与维护负担，但无法替代多指手。需要同时按压防童锁并扭盖、指甲式挑起薄边、较大力矩或较高负载的任务仍受机构约束。附录的硬件对比价格均属当时估计，能力对照以公开演示为依据，不应解释成当前商业报价或标准化性能排名。

S026 · 英文 PDF p.16

论文公开两类未学会的行为。拆糖纸用 50 条演示，初步 10 次测试中抓取 10/10、两端拉动 8/10、最终剥开 0/10；另一个允许每颗糖最多 10 次尝试的协议在 5 颗中成功 3 颗。这两个结果的重试预算不同，必须并列说明。平放的小密封袋也能抓起，却难以完成后续三步空中重抓和拉开。

S027 · 英文 PDF p.16

失败机制集中在细小接缝的可辨识性、变形后的几何变化与数据不足。预训练、更多数据和更强视觉被提出为后续方向，而不是已验证修复。对工程部署，先判断失败是“看不见”“看见但未建模”“动作不可达”还是“控制跟踪失败”，才能选择相应改动。

12. 连接扩散策略、流匹配与事件观测

S028 · 英文 PDF p.4, 5, 6

> 【读者推演】ACT 和 Diffusion Policy 都用动作块，但默认推理的随机性、生成目标与融合方式不同。ACT 以固定潜变量一次输出动作块，再融合重叠块；扩散策略从噪声迭代生成动作块，再滚动执行前缀。比较 lerobot-flow-matching 时，可以先统一动作块长度、观测、低层控制和执行时域，再比较 CVAE、噪声回归与速度回归。

S029 · 英文 PDF p.3, 5, 6

> 【读者推演】50 Hz 动作更新只有 20 ms 时间预算，论文报告约 10 ms 推理并不意味着整个闭环还有 10 ms 稳定余量。四路相机的最新帧可能已有几十毫秒年龄，USB 传输、编码、排队和执行也需计入。最有用的日志是每个动作携带其最后观测时间与生效时间，而不是仅记录 GPU 平均前向耗时。

S030 · 英文 PDF p.3, 6, 16

> 【读者推演】EEPTV 的事件视觉经验可用于高动态与低延迟观测，但透明低对比目标未必产生足够事件。应先测目标区域的事件率、遮挡、运动与照明变化，再决定事件是否能改善可观测性。不能仅凭事件相机微秒级时间戳，就声称动作策略拥有微秒级闭环或解决静止透明边缘。

统一术语与底本

English	中文	使用边界
ALOHA	低成本开源双臂遥操作系统	硬件与采集接口
ACT	基于 Transformer 的动作分块	策略学习方法
leader / follower	主导臂 / 跟随臂	命令源 / 实际执行与观测
action chunk	动作块	固定 k 个连续目标动作
temporal ensembling	时间集成	同一执行时刻的重叠预测融合
CVAE	条件变分自编码器	训练时推断潜变量，测试默认 z=0
style variable	风格变量	z，不等于任务 ID
absolute joint target	绝对关节目标角	不是增量角，也不是力矩
reconstruction loss	重建损失	正文 L1，算法伪码 MSE 存在差异
success rate	成功率	百分比与百分点须分开

官方底本：https://arxiv.org/abs/2304.13705v1
本稿页码定位以随附 PDF 为准；原文中的时间语境属于论文版本日期。

构建：本地 KaTeX 静态公式 + Microsoft Edge A4 打印。中文内容与中英对照稿来自同一份 content.json。