扩散策略：从动作分布到可执行的视觉闭环

Diffusion Policy: Visuomotor Policy Learning via Action Diffusion

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, Shuran Song

RSS 2023 工作的扩展期刊版本（arXiv v5）
底本 arXiv:2303.04137v5 · 2024-03-14 · 原文 19 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：动作分布不能被一条均值轨迹替代
2. 2. 去噪学习：一次训练样本实际包含什么
3. 3. 条件生成与能量策略：为什么无需估计配分函数
4. 4. 三种时域与闭环预算
5. 5. 卷积与 Transformer：生成头的取舍
6. 6. 视觉编码、归一化与采样器
7. 7. 模态一致性、位置控制与停顿
8. 8. 与线性控制的联系：一个可检验的极限
9. 9. 实验协议：哪些数字能够公平比较
10. 10. 核心数值：强项也体现在困难子任务上
11. 11. 真机闭环：末端成功、扰动恢复与统计口径
12. 12. 扩展版双臂任务与硬件控制栈
13. 13. 局限与结论：表达能力不是安全保证
14. 14. 连接 lerobot-flow-matching 与 EEPTV 的复现路线

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：动作分布不能被一条均值轨迹替代

S001 · 英文 PDF p.1, 2

这篇工作将行为克隆的输出对象从“当前动作”改成“未来动作序列的条件分布”。输入仍是视觉与本体感觉，学习目标仍来自专家演示；改变的是策略的表达方式与执行节奏。对一个障碍物，左绕与右绕都可能正确，但两条路径的均值可能正好穿过障碍。模型因此需要保留多个合法解，而不是只降低逐点均方误差。

S002 · 英文 PDF p.1, 2

原文 v5 的正文和 PDF 摘要写的是 15 项任务、4 组基准；arXiv 摘要网页仍写 12 项。这份阅读稿以随附的 19 页 v5 PDF 为准。46.9% 是按任务计算的相对提升再平均，不是成功率增加 46.9 个百分点，也不是所有真实机器人任务的平均成功率。

2. 去噪学习：一次训练样本实际包含什么

S003 · 英文 PDF p.2, 3

设观测窗口为 O_t，干净动作块为 A_t^0，扩散层级为 k。训练时先抽一个演示窗口，再抽一个噪声层级与一份高斯噪声；网络根据观测、带噪动作块和层级预测噪声。扩散层级是生成算法内部的时间，不是机器人运动的物理时间。两个下标一旦混淆，就会误把“十次去噪”理解为“机器人执行十步”。

S004 · 英文 PDF p.2, 3

为便于核对噪声尺度，下面采用标准 DDPM 写法重述训练目标。$\bar{\alpha}_k$ 是累计信号保留系数，ϵ 为单位高斯噪声。原文式 (3)、(5) 采用更压缩的加噪记法，不能据此省略实现中与噪声调度相配套的缩放。最优均方误差预测器给出条件均值 $\mathbb{E}[\epsilon \mid A_t^k, O_t, k]$；这并不意味着最后的动作退化为无条件均值，因为采样会保留不同初始化所到达的模式。

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

$$\mathcal{L}_\epsilon = \mathbb{E}_{(O_t, A_t^0), k, \epsilon}\left[\left\|\epsilon - \epsilon_\theta(O_t, A_t^k, k)\right\|_2^2\right]$$

3. 条件生成与能量策略：为什么无需估计配分函数

S005 · 英文 PDF p.3, 5, 6

策略只生成动作而不同时重建观测。也就是说，目标是 $p(A_t \mid O_t)$，不是 $p(A_t, O_t)$。相机特征可在一次规划开始时提取并复用，避免每个去噪步骤都重跑视觉骨干；动作维度也远小于原始图像维度。这个计算图拆分是实现实时性的关键，不能把系统速度全部归功于扩散采样器。

S006 · 英文 PDF p.5, 6

若隐式策略写为 $p_\theta(a \mid o) = \exp[-E_\theta(o,a)]/Z(o,\theta)$，对动作求导时，$Z$ 的导数为零，因为它不依赖被积分掉的动作变量。得分函数由此避开配分函数的直接估计。原文借此解释其相对 IBC 的训练稳定性；但噪声预测等于得分函数的说法只在忽略符号和噪声尺度时成立，标准对应关系还需要除以 $\sqrt{1-\bar{\alpha}_k}$。

S007 · 英文 PDF p.6

稳定性证据来自训练动作误差与 rollout 成功率的曲线，不是由一个损失值单独推出。能量损失下降而推理动作不改善，说明训练目标、负采样和动作搜索之间可能发生错配。部署比较应同时检查策略 rollout 与 checkpoint 选择成本，避免给一种方法无限筛选机会而只给另一种方法最后一个权重。

$$\nabla_a \log p_\theta(a \mid o) = -\nabla_a E_\theta(o,a)$$

$$\nabla_{A_t^k} \log p_k(A_t^k \mid O_t) \approx -\frac{\epsilon_\theta(O_t, A_t^k, k)}{\sqrt{1-\bar{\alpha}_k}}$$

4. 三种时域与闭环预算

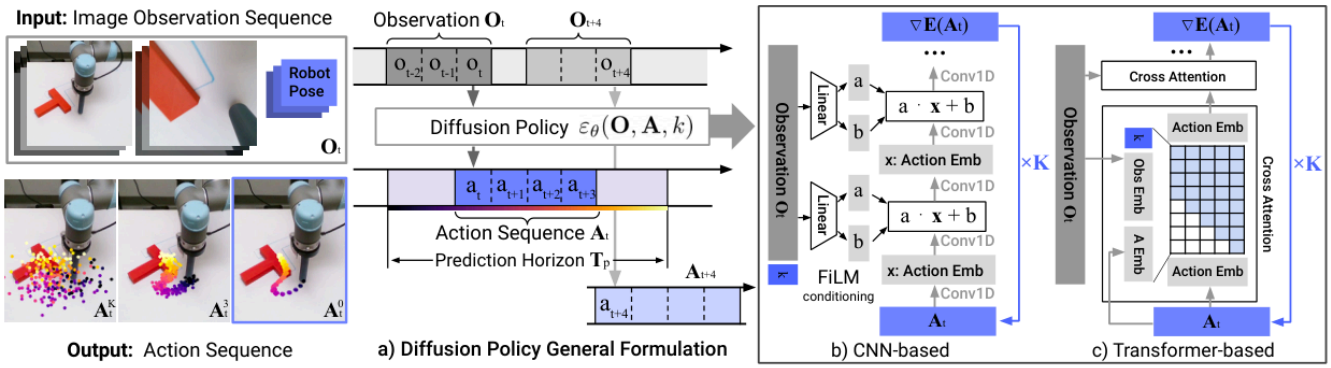
S008 · 英文 PDF p.3, 5, 8

论文区分观测长度 T_o、预测长度 T_p 和执行长度 T_a。一次预测覆盖较长的未来，只执行其前部，然后用新观测再规划。较长 T_p 帮助保持动作间的相关性；较短 T_a 提高对外部扰动的反应能力。它们不是同一超参数，动作序列预测也不等于开环播放整条轨迹。

S009 · 英文 PDF p.5, 8

在多数消融任务中，执行 8 步是较好的折中；这个结果依赖任务与控制频率。以 10 Hz 动作采样为例，8 步对应 0.8 s，而同样的步数在 50 Hz 下只有 0.16 s。工程上应先用秒定义允许的感知更新间隔，再换算步数。论文的模拟时延实验显示部分设置可容忍 4 步延迟，但这不能替代对相机年龄、网络抖动与接触安全的实际测量。

$$T_{\text{refresh}} \simeq T_a / f_{\text{cmd}}, \qquad T_{\text{latency}} = T_{\text{camera}} + T_{\text{vision}} + T_{\text{denoise}} + T_{\text{transport}}$$



原图 2：条件动作生成与两种生成头 · 英文 PDF p.3

从左向右看：观测只在规划开始时编码；动作噪声在内部迭代中逐步去除。右侧是替代架构，不是串联的两个阶段。图中机器人时间 t 与去噪层级 k 必须分开。

5. 卷积与 Transformer：生成头的取舍

S010 · 英文 PDF p.3, 4

卷积分支用一维时序网络处理动作序列，并用 FiLM 将观测与扩散层级注入各层。对通道特征做 $\gamma(O_t, k) \odot h + \beta(O_t, k)$ 变换，可在不把相机帧反复拼入动作张量的情况下控制去噪。原文推荐新任务先试这一分支，因为多任务间超参数较一致。

S011 · 英文 PDF p.3, 4

Transformer 分支把带噪动作看作 token，把观测作为交叉注意力的上下文。原图还给出动作 token 的因果掩码。它不是逐 token 自回归地产生动作：一个去噪调用仍同时输出动作块对应的噪声预测。其优势主要出现在动作变化较快、卷积平滑先验不合适的设置；代价是调参更敏感，增大网络深度也不总能提高表现。

6. 视觉编码、归一化与采样器

S012 · 英文 PDF p.4, 16, 17

默认视觉骨干为 ResNet-18，采用空间 softmax 代替全局平均池化以保留位置信息，以 GroupNorm 代替 BatchNorm 配合 EMA 训练。不同视角有独立编码器，各时刻独立编码后拼接。训练使用随机裁剪，推理使用相同尺寸的中心裁剪。这些细节直接改变像素位置与动作目标之间的对应关系。

S013 · 英文 PDF p.4, 16, 17

对大多数非旋转动作维度，附录使用逐维最小最大值映射至 $[-1, 1]$；近常量维度只平移，避免极小范围带来的放大。位置控制的旋转使用连续 6D 表示，速度控制则使用轴角。不能把任意 z-score 标准化直接与每步裁剪到 $[-1, 1]$ 的采样代码拼接，否则训练分布尾部可能永远无法生成。

S014 · 英文 PDF p.4, 16, 17

原文第 3.4 节给出 100 个训练层级、10 次 DDIM 推理、RTX 3080 上约 0.1 s 的示例；附录 A.4 与表 7、8 的真实任务则记录 16 次推理。这里保留两种出处，不把 10 次统一改写成实验最终配置。DDIM 减少的是生成迭代数，不会自动消除视觉、队列和底层执行时延。

设置	观测 / 执行 / 预测步数	训练 / 推理去噪步数
CNN：多数模拟任务	2 / 8 / 16	100 / 100
CNN：真实 Push-T	2 / 6 / 16	100 / 16
CNN：真实浇酱、抹酱、翻杯	2 / 8 / 16	100 / 16
Transformer：真实 Push-T	2 / 6 / 16	100 / 16

7. 模态一致性、位置控制与停顿

S015 · 英文 PDF p.4, 5

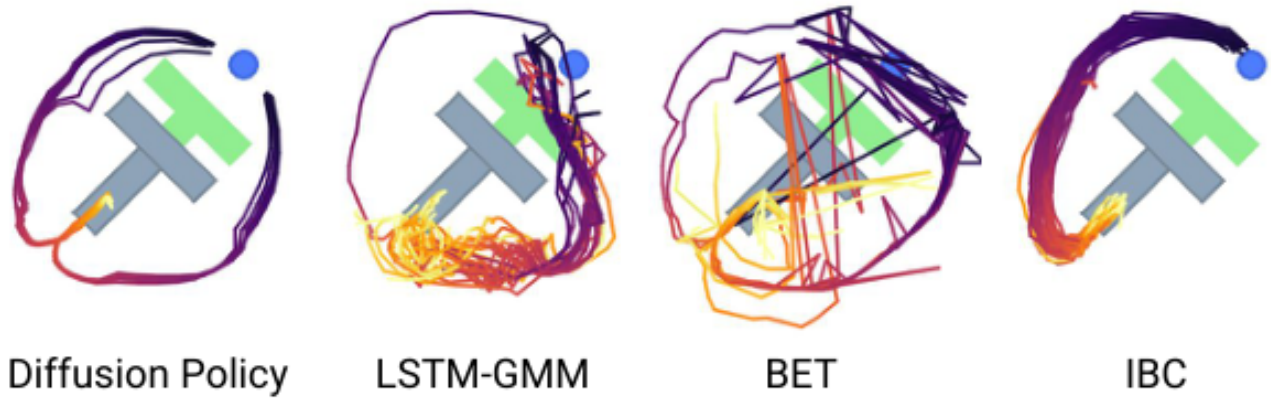
多模态有两个层次：当前接触点可从左右两侧接近，是短期多模态；厨房子任务或双块推动可按不同顺序完成，是长期多模态。能够在样本集合中出现两种模式，只是第一步；同一次 rollout 还必须坚持一个可行方案。动作块把跨时间相关性纳入联合预测，减少逐步采样在模式间来回跳变。

S016 · 英文 PDF p.5, 8

位置控制在本文的扩散策略实验中优于速度控制。作者的解释是：策略能够承担位置目标更明显的多模态，同时避免速度积分导致的累积误差。这是受任务与控制器条件约束的经验结论，不是位置控制在所有机器人上的普遍优越性。相对地，基准方法在报告中使用各自较优的速度动作空间，因此对比的是完整策略配置。

S017 · 英文 PDF p.5

演示停顿不一定是垃圾数据：等待粘稠酱汁装满勺子就是有效的任务阶段。单步策略若把停顿误学为永久停止，rollout 会卡住。动作块能把“先停一下再继续”作为时间结构建模。清理数据时应区分操作员分心、任务必要等待和安全停止，而不是统一删除零速度片段。



原图 3：左右绕行的模态与轨迹一致性 · 英文 PDF p.5
比较的不是单个动作误差，而是 40 步 rollout 是否选定并保持一侧路径；高频切换模式会形成无法执行的折中。

8. 与线性控制的联系：一个可检验的极限

S018 · 英文 PDF p.6

扩展版用线性系统与线性反馈专家作一致性检查。若状态满足 $s_{t+1} = As_t + Ba_t + w_t$，专家为 $a_t = -Ks_t$，那么单步动作分布的中心就在这个反馈律上。多步动作预测则隐含对闭环矩阵 $A - BK$ 的理解。它揭示“纯行为克隆”在预测未来动作时也可能学习任务相关动力学，而非证明网络具备完整物理世界模型。

S019 · 英文 PDF p.6

下式给出零均值过程噪声下的条件期望。非线性系统中，未来状态的不确定性经过策略映射后可能形成多个动作模式，不能简单把期望穿过非线性函数。原文对最优去噪器的 σ_k 记法同时涉及方差与尺度的表述，复现时应回到实际 scheduler 的定义，而不是直接照抄一个比例常数。

$$\mathbb{E}[a_{t+\tau} \mid s_t] = -K(A - BK)^{\tau} s_t$$

9. 实验协议：哪些数字能够公平比较

S020 · 英文 PDF p.6, 7, 8, 17

模拟实验包括 Robomimic、Push-T、双块推动和 Franka Kitchen。Robomimic 同时区分熟练单人数据 PH 与混合多人数据 MH、状态与图像观测。大多数任务报告成功率，Push-T 用目标区域覆盖度；双块推动和 Kitchen 还按完成子任务数分别统计。不能把这些指标都当作完全相同的二元成功率。

S021 · 英文 PDF p.7, 8

表 1、2 使用“最佳 checkpoint / 最后 10 个 checkpoint 平均”格式，两者又跨 3 个训练种子平均。正文协议写 50 个环境初始条件，但第 7 页脚注明确：由于评估代码错误，Robomimic 实际只用了 22 个初始条件，并对基线采用了相同处理。阅读结果必须保留这条修正，而不是继续用名义 1500 次 rollout 描述每个 Robomimic 数值。

S022 · 英文 PDF p.17

附录先在每个任务选择最优基线与最优扩散分支，再计算相对提升并平均，且忽略 MH 列。这个聚合显示方法族的潜力，但会结合不同任务上的不同分支，不能解释成单一固定网络在所有任务上的统一提升。Kitchen 数据量也有内部差异：表 3 写 656 条，正文写 566 条；本稿不擅自选定一个值作为训练事实。

$$r_i = \frac{\max_{v \in \mathrm{DP}} S_{i,v} - \max_{b \in \mathrm{baseline}} S_{i,b}}{\max_{b \in \mathrm{baseline}} S_{i,b}}, \qquad \bar{r} = \frac{1}{N}\sum_i r_i = 0.46858$$

10. 核心数值：强项也体现在困难子任务上

S023 · 英文 PDF p.7, 8, 9

视觉 Transport 的 PH 设置中，CNN 扩散策略的最佳/末段平均为 1.00/0.93，LSTM-GMM 为 0.88/0.62；ToolHang 为 0.95/0.73 对 0.68/0.49。这些差异比简单 Lift 的接近饱和结果更能揭示高维序列建模的价值。双块推动则是 Transformer 明显强于 CNN 的反例，提醒我们不能把“卷积更稳”误写成“卷积总是更好”。

S024 · 英文 PDF p.8, 9

视觉编码消融在 Square PH 上进行。CLIP ViT-B/16 从零训练只有 0.22，冻结预训练为 0.70，微调可达 0.98。ResNet-18 从零训练为 0.94，冻结为 0.58，微调为 0.92。这组结果支持“视觉表征需要适配动作任务”，但并不支持“预训练无用”或“更大骨干必然更强”。

任务 / 指标	基线	CNN 扩散	Transformer 扩散
视觉 Transport PH：最佳 / 末段	0.88 / 0.62（LSTM）	1.00 / 0.93	0.98 / 0.81
视觉 ToolHang：最佳 / 末段	0.68 / 0.49（LSTM）	0.95 / 0.73	0.76 / 0.47
BlockPush p2	0.71（BET）	0.11	0.94
Kitchen p4	0.44（BET）	0.99	0.96

11. 真机闭环：末端成功、扰动恢复与统计口径

S025 · 英文 PDF p.9, 10, 11, 17, 18

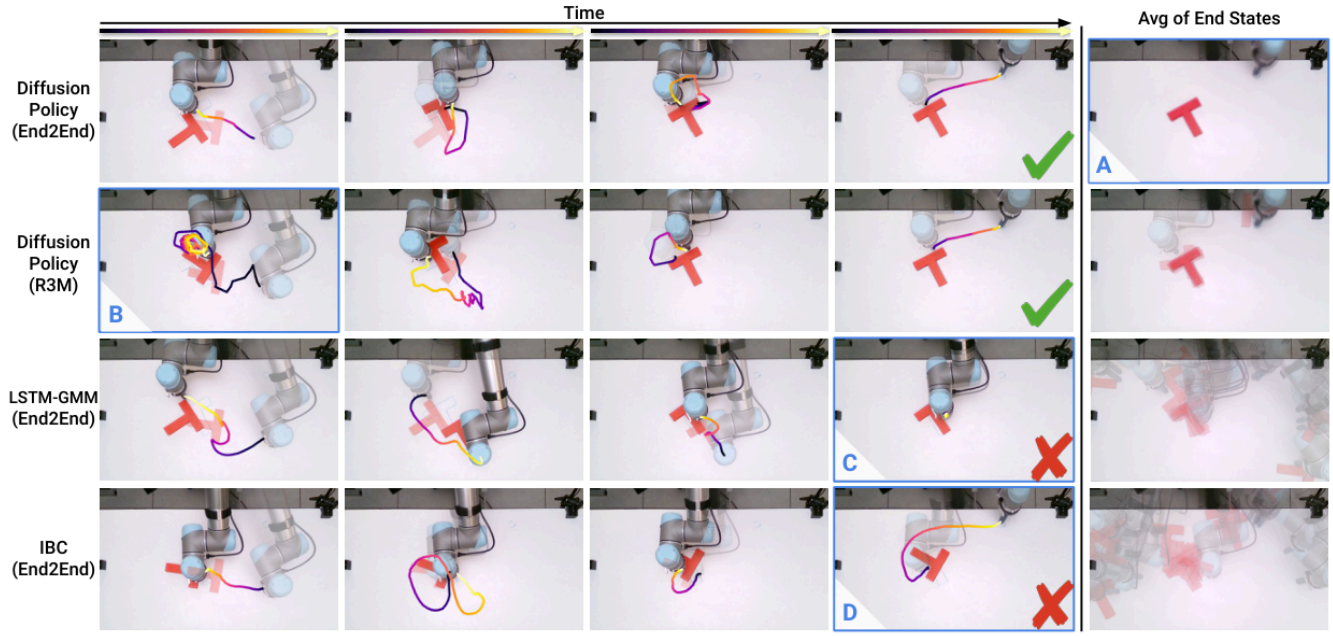
真实 Push-T 不是模拟任务的直接复刻：机器人需先精确放正 T 块，再移到终点区域避免遮挡，指标在最后状态计算。20 个初始条件相同的测试中，端到端 CNN 扩散策略成功率 0.95、平均 IoU 0.80；人类为 1.00、0.84。任务按末端在终点区停留 0.5 s 或到达 60 s 超时终止。

S026 · 英文 PDF p.9, 10

遮挡、推动中挪动物块、结束时再次挪动物块是独立的扰动测试，不是 20 次标准评估中混入的随机条件。示例显示策略能重新进入推动阶段；它证明了所展示场景的闭环反应能力，不提供所有扰动分布下的鲁棒概率。翻杯任务则要求杯口向下且把手朝左，20 次测试成功率 0.90，基准 LSTM-GMM 在这些条件下未成功抓取。

S027 · 英文 PDF p.10, 11, 17, 18

浇酱和抹酱分别需要等待粘稠液体、维持周期运动并按视觉覆盖反馈修正。结果表中浇酱 IoU 为 0.74（人类 0.79），抹酱覆盖率为 0.77（人类 0.79）。表 3 给出 90 条演示，而附录 C.2 写每项收集 50 条、90% 用于训练；来源存在差异，复现实验时应核对公开数据而不是将两种数字合并。



原图 7：真实 Push-T 的完整阶段对照 · 英文 PDF p.10

列为关键时刻，行是方法。末列的叠加图显示末态一致性；不要只看接近目标时的中间帧。

12. 扩展版双臂任务与硬件控制栈

S028 · 英文 PDF p.11, 12, 18

双臂输入扩展为两端末端位姿、夹爪宽度、实际值和期望值，并加入两张场景图与两张腕部图。打蛋器任务用触觉遥操作采集数据；操作者无触觉的 10 次尝试全部失败，有触觉的 10 次全部成功。这个对照首先说明数据采集接口影响示范可获得性，不能把它归为策略网络能力的单独提升。

S029 · 英文 PDF p.12, 18, 19

垫子展开任务用 162 条演示，在 20 次测试中成功率 0.75；叠衫用 284 条演示，同样为 0.75。失误集中在最初抓取、重复失败动作以及结束阶段不停整理。低层控制并未被端到端网络取代：附录的差分运动学 QP 处理关节限制、桌面和双臂避碰，末端命令还要插值。触觉采集的力矩控制模式与策略推理所用的非触觉控制模式不同。

S030 · 英文 PDF p.18

UR5 示例中，10 Hz 的上层末端目标被插值为 125 Hz 命令，速度限制为 0.43 m/s，并限制在桌面上方区域。Franka 站也用高频中层控制器跟踪 10 Hz 遥操作或策略命令。因此“策略 10 Hz”与“电机控制 10 Hz”完全不同；评估实时性必须指明测量的是哪一层。

13. 局限与结论：表达能力不是安全保证

S031 · 英文 PDF p.12, 13

本工作仍是行为克隆：示范不足、分布外物体或错误恢复数据缺失时，生成模型不会自动获得任务成功奖励。模型能表达复杂分布，也可能精确地复现错误、偏见和不充分覆盖。更丰富的负样本、强化学习或在线修正属于后续训练范式，而非这套监督目标已经实现的功能。

S032 · 英文 PDF p.13

计算成本是另一条明确边界。动作块减少重复规划需求，但高频、强接触任务仍可能被去噪延迟限制。本文的结论应读成：在所测试的数据、控制器和评估条件下，策略表示是重要瓶颈，动作扩散与闭环序列执行能缓解它。它没有给出形式化的稳定性、安全性或硬约束可行性保证。

14. 连接 lerobot-flow-matching 与 EEPTV 的复现路线

S033 · 英文 PDF p.3, 4, 16, 17

> 【读者推演】在 lerobot-flow-matching 中，优先保持数据窗口、动作规范化、视觉骨干、控制器和评估初始条件不变，只替换动作生成目标与采样器。DDPM 噪声回归和条件流匹配速度回归使用的时间方向、噪声端点与输出尺度不同，不能只将变量名 epsilon 改为 velocity。记录每次规划的视觉耗时、每次生成头耗时、NFE、总延迟与成功率，才能区分算法进步和系统缓存收益。

S034 · 英文 PDF p.3, 8, 18

> 【读者推演】EEPTV 的事件-帧融合可以作为低延迟观测编码研究，但本文没有使用事件相机。首先应保留事件原始时间戳、曝光时间、动作生效时间和坐标变换，定义观测年龄与窗口截止时刻。用等时窗事件堆叠替代 RGB 后若看到性能改善，还需对照更高帧率图像或等延迟输入，避免把采样频率和表征能力混为一谈。

S035 · 英文 PDF p.5, 8, 13

> 【读者推演】推荐实验顺序是：先跑通已知动作块基线；再在相同 NFE 下比较目标；再固定真实延迟预算减少步数；最后施加可重复的物块扰动和传感延迟。每一阶段同时报告末段平均与最佳 checkpoint、失败类型和动作约束违规次数。单个漂亮视频不足以支持“20 ms 实时闭环”的工程结论。

统一术语与底本

English	中文	使用边界
action chunk	动作块	联合生成的连续动作序列
observation horizon	观测窗口长度	T_o
prediction horizon	预测窗口长度	T_p
execution horizon	执行窗口长度	T_a
receding-horizon control	滚动时域控制	执行前缀后重新规划
score function	得分函数	对样本求导的对数密度梯度
noise predictor	噪声预测器	ϵ_θ，需带噪声尺度理解
NFE	函数求值次数	一次生成中的网络调用数
PH / MH	熟练单人 / 混合多人演示	不要合并统计口径
spatial softmax	空间 softmax	保留位置特征的池化

官方底本：https://arxiv.org/abs/2303.04137v5
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。