arXiv:2611.28732v1 [cs.RO] 21 Nov 2026

# Entangle：扩展跨未见任务和上下文的人类视频上下文世界动作模型

**Chendong Xiang***, **Chengwen Zhang***, **Ruowen Zhao**, **Zeyuan Wang**, **Shuhe Huang Rongxu Cui**, **Min Zhao**, **Fan Bao**, **Hengkai Tan**, **Jun Zhu**

清华大学、GensPI Technology、MotuBrain

**摘要**

上下文学习（ICL）使基础模型能够仅凭上下文执行未见任务，而无需更新参数。在机器人操作中，人类视频提供了指定任务执行方式的上下文。然而，在世界-动作模型（WAM）上扩展人类视频上下文学习面临两项挑战：任务层面的人类–机器人配对与机器人动作在时间上并未对齐，而且标准动作损失会使策略依赖初始场景或任务身份捷径，而不是跟踪上下文。我们提出 **Entangle**，一种上下文世界-动作模型，可将单个人类视频和当前机器人观测映射为未来视觉状态与可执行动作。为扩展时间对齐的训练数据，我们通过视频到视频生成将记录的机器人视频转换为匹配的人类上下文，并辅以真实配对记录。在模型训练中，我们将可变长度的视频窗口与任务内对比目标相结合，使模型能够区分配对的动作轨迹与同一任务的其他执行过程，从而避免任务身份捷径。在不进行任务特定微调的情况下，随着对齐训练数据规模扩大，Entangle 的性能持续提升：在 Astribot S1 机器人上的六个未见任务中达到平均 83.5/100 的得分，并在七个未见 RoboTwin任务上取得 69.5% 的平均成功率，较 Zero-WAM 高出 22.6 个百分点。在上下文分布外场景下，尽管仅使用场景对齐的配对数据进行训练，Entangle 仍能将上下文指定的行为迁移到重新排列的布局和替换后的工具上。

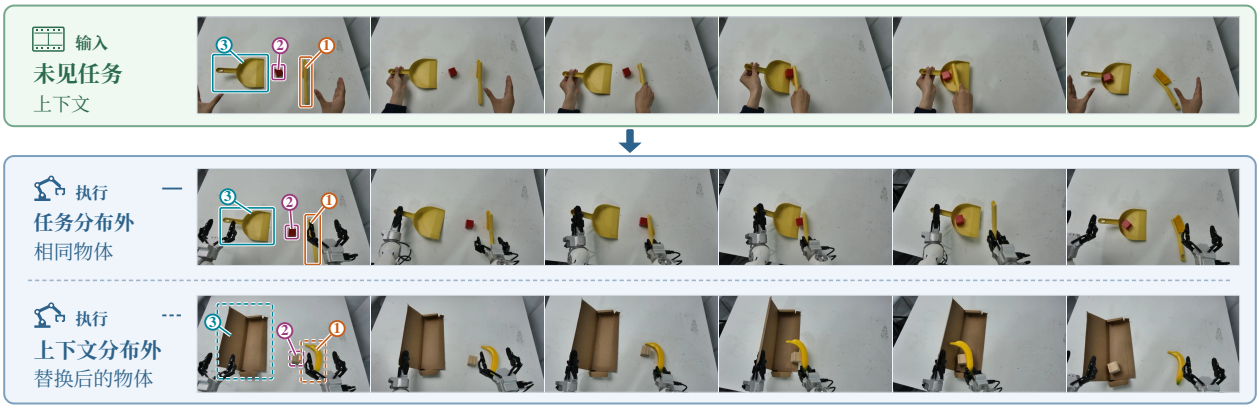



**图 1 执行未见任务**。输入上下文（上）展示了人类使用刷子（1）将塑料块（2）扫入簸箕（3）中，这项任务不在训练集中。**任务分布外**（中）：以此上下文为条件，机器人使用相同物体执行未见任务。**上下文分布外**（下）：机器人使用与输入上下文中不同的物体执行相同的交互，即用香蕉（1）将木块（2）扫入纸箱（3）中；虚线轮廓标示替换后的物体。

*这些作者贡献相同，并列为共同第一作者。

## 1 引言

上下文学习（ICL）使基础模型能够仅凭上下文执行未见任务，而无需更新模型参数 [1]。在机器人操作中，语言指令通常无法充分说明多步骤顺序、空间约束和细粒度运动，而单个人类示范视频则能直接展示任务应如何执行 [2–6]。世界-动作模型（WAM） [7–11], 在共享生成式骨干网络中联合建模未来视觉动态与机器人控制，为此类视频条件策略提供了自然的基础：给定一段单个人类示范视频作为上下文，WAM 无需更新参数即可为未见操作任务预测未来视觉状态和可执行动作。

然而，在 WAM 中扩展人类视频上下文学习面临两项核心挑战。首先，人类视频与机器人视频通常仅在任务层面配对。无论是独立采集，还是通过文本到视频生成从编辑后的初始帧合成 [9, 12], ，任务层面的配对都共享高层目标，但在运动路径、接触时序和执行速度上存在差异。如果人类视频与机器人真实动作之间没有时间对齐，模型就无法定位示范中的哪个片段对应目标动作片段，从而使视频提示退化为粗略的任务标签。其次，即使配对轨迹已对齐，标准的动作预测损失仍是对每条轨迹单独进行监督，因此存在走捷径的空间：当某个任务存在多种有效执行方式（例如不同的操作顺序或避障路径）时，策略可以通过推断任务身份，或依赖初始场景布局，而非跟踪示范所呈现的具体进程，来降低损失。

为应对这些挑战，我们提出**Entangle**，这是一种上下文世界-动作模型，可将单个人类示范视频和当前机器人观测映射为未来的视觉状态与可执行动作。首先，为扩展时间对齐的训练数据，我们通过参考条件视频到视频转换（V2V）生成，将录制的机器人视频转换为匹配的人类示范，并辅以真实配对记录（图2）。与先编辑首帧、再通过文本驱动生成的方式不同，V2V 转换保留了摄像机视角、物体交互以及源机器人轨迹的时序，因此生成的每段人类视频都与录制的真实动作保持时间对齐。其次，我们将人类示范和当前机器人观测编码到共享的视频-动作骨干网络中，并沿时间轴拼接它们的令牌（图3）。在模型训练中，我们结合可变长度的视频窗口与任务内 InfoNCE 对比目标 [13] ，其灵感来自负样本感知微调 [14]：通过拉近预测动作与视频配对轨迹的距离，同时推远同一 任务中的其他轨迹，该方法能够防止任务身份捷径，并迫使策略遵循示范中的运动。

我们在 Astribot S1 机器人和 RoboTwin 基准上评估 Entangle [15, 16]。在无需任务特定微调的情况下，随着对齐训练数据规模的扩大，Entangle 的性能持续提升：在 Astribot S1 上的六个未见任务中，平均得分达到 83.5/100；在七个未见的 RoboTwin任务上，平均成功率达到 69.5%，优于 Zero-WAM [9] 22.6 个百分点。更令人惊讶的是，尽管我们的训练对严格遵循场景对齐，Entangle 不仅能处理具有相同物体的未见任务（图1，中）：它还能将示范行为迁移到重新排列的布局和替换后的工具上，例如，从刷子和簸箕示范中学习使用香蕉将木块清扫进纸箱（图1，上、下）。

我们的主要贡献如下：

- **可扩展的时间对齐数据：** 我们通过参考条件视频到视频生成，将记录的机器人轨迹转换为时间匹配的人类演示，使对齐能够随机器人视频语料库扩展，而无需人工复演。
- **上下文内 WAM 训练：** 我们沿时间维度拼接示范与机器人视频令牌，在可变长度窗口上进行训练，并加入一种任务内 InfoNCE 损失，以消除任务身份捷径。
- **数据扩展与跨场景迁移：** 未见任务得分随对齐数据的增加而单调提升；基于场景对齐的训练可迁移至新的布局和工具。

## 2 相关工作

### 2.1 机器人操作中的上下文学习

上下文学习使冻结的机器人策略能够从部署演示中推断任务，例如 ICRT 中的感觉运动轨迹 [3], 、Instant Policy 中的演示图 [4], , 或 RICL 中检索到的片段 [5], ; 相关方法通过任务嵌入 [17, 18], 、帧注意力 [19, 20]，或技能先验 [21–23]扩展单次策略。为利用人类视频，先前工作学习视觉特征 [24], 、潜在动作 [25], 或游戏计划 [26], ; 在 EgoMimic 中使用第一人称数据进行联合训练 [27], ; 在 RHyME [28] 和 MimicDroid [29], 中检索游戏片段；或通过 ViP 中的视频检索 [30], 、BC-Z 中的语言落地 [31], 以及 Vid2Robot 中的交叉注意力 [2], 以单个人类视频作为部署条件。然而，将示范简化为任务标识符，会使策略局限于回忆训练行为 [32], ; 而要泛化至未见过的流程和新颖的物体用途，则需要追踪交互是如何展开的，这方面的探索包括子目标或轨迹预测 [33–35], 、时间对齐 [36–39], 、JEDI 中的任务内关键帧分离 [40], 、SimpleICL 中的布局变化 [41], 以及 The Imitator Game 中的功能替代 [42]。Entangle 使用单个人类视频，在世界-动作模型中指定流程，从而泛化至未见过的流程（任务分布外）以及变化的场景或工具（上下文分布外）。

### 2.2 世界-动作模型

世界-动作模型联合预测未来观测和动作 [10, 43–46], , 扩展视觉预见 [47], 、从视觉计划中进行动作恢复 [48–50], 、用于控制的预测性表征 [51–53], , 以及适用于策略学习 [7, 11, 54–56] 或展开合成 [57] 的视频生成器。近期的 WAM 进一步融入上下文和示范：ICWM [58] 通过自交互推断系统属性，WATCH [59] 在任务配对的人类视频上进行预训练，LingBot-VA 2.0 [8] 和 WorldScape Policy 2.0 [60] 接受演示提示，而 Zero-WAM [9] 生成任务匹配的人类视频片段，同时预测时间间隔更大的未来片段，以促进提示使用。然而，仅以示范作为条件，并不能确保预测的动作会依据任务级线索遵循示范中的执行过程；Entangle 通过将时间对齐训练与任务内动作对比相结合来解决这一问题 [13] ，从而区分同一任务的不同执行方式。

### 2.3 测试时训练

元模仿根据示范 [61, 62], 调整策略参数，而测试时训练则根据测试输入 [63–65]调整模型或快速权重。WAM-TTT [66] 在展开前根据人类视频更新快速权重，而 RoboTTT [67] 则随着机器人的视觉运动历史不断增长而更新权重。Entangle 在部署时保持权重固定，针对每次预测直接以人工演示和当前机器人观测为条件。

### 2.4 运动重定向与配对数据

运动重定向将物体相对运动 [68, 69] 或人手轨迹 [70–72] 迁移到机器人上；BEHAVE [73]记录人-物接触，CORE4D [74]则将双人重排运动重定向到新的物体形状。机器人训练数据还可以来自人机配对记录 [75–77] ，也可以通过人类视频素材合成机器人示范 [78–81], ; 而 Gen2Act [82] 则根据初始图像和任务描述生成一段人类视频。Entangle 将完整的机器人记录用作运动参考，验证生成的人类视频与来源在时间和交互一致性方面是否相符，并将这些片段与同步人类重演相结合，使每个采样的示范窗口都能与记录的机器人动作直接配对。

## 3 Ent-Dataset 数据集构建

### 3.1 数据集概述与严格配对

Ent-Dataset 中的每个训练样本都将一段人类示范视频与一条录制的机器人轨迹配对，其中包含相机观测、本体感知状态和动作（图2（c））。如图2（b）所示，该数据集包含 21.3k 组配对数据，划分自互不重叠的机器人回合，用于训练和验证；其中包括来自 Astribot S1 的 5.3k 组真实配对数据（**Ent-Real**）以及来自 Astribot 和 AgiBot 的 16.0k 组合成配对数据（**Ent-Syn**）。




**图 2 Ent-Dataset 概览。**(a) **Ent-Syn 流程：**源机器人轨迹依次经过 (1) 关键帧手动编辑、(2) 运动引导视频生成和 (3) 视觉语言模型验证，以确保视觉质量和动作一致性；随后将生成的人类视频与原始机器人观测和动作配对。(b) **数据集构成：**同心圆环（由内向外）依次展示全部 21,269 个配对回合按数据来源（Ent-Real、Astribot、AgiBot）、人类视频来源（人类、Seedance、Wan）、训练集/验证集划分（19,115 / 2,154）和任务类别（635 个任务）的细分情况。(c) **配对示范：**展示 Ent-Real 和 Ent-Syn 子集中的示例，涵盖五个匹配的时间戳，呈现时间对齐的机器人与人类执行过程。

为了根据人类视频监督细粒度动作预测，每个目标动作片段都需要在示范中有一个局部视觉参照。当人类视频和机器人视频仅按任务标签配对，或由编辑后的初始帧合成时， [9], 后续运动路径和接触时序仍与机器人真实动作未对齐。为提供明确无歧义的监督，我们采用严格配对：以录制的机器人轨迹为锚点，配对的人类视频需匹配（i）摄像机视角和场景布局，（ii）操作物体及交互事件序列（例如伸手、抓取、搬运和释放），以及（iii）每个事件的时序。这种对齐要求策略追踪视觉进程，而非依赖初始场景或任务身份捷径；我们在表2中对严格配对进行消融，并在第5.6节评估部署时的跨场景迁移。

### 3.2 Ent-Real

我们通过让人类在同一工作空间中、以与机器人头部摄像头相匹配的视角，重现录制的 Astribot S1 轨迹，从而构建真实配对数据。录制过程中，演示者观看源机器人视频回放，并同步进行操作，使物体交互、运动路径和执行节奏保持一致。

### 3.3 Ent-Syn

为了扩展配对语料库而无需逐回合进行人工复演，我们通过三阶段流程将现有机器人视频转换为人类演示（图2(a)）。

**关键帧编辑。**W　　从机器人视频标注中提取原子动作描述（包括 Qwen3.8　　-

生成的描述），并抽取来源关键帧进行外观编辑。使用 HaGRID [83]中的人手图像作为外观参考，*gpt-image-2* 将画面中可见的机器人手臂和夹爪替换为人手和前臂，同时保留场景、任务物体、视角和接触关系。编辑后的帧通过视觉质量检查后，再进行视频生成。

**视频生成**。使用*Seedance 2.5* 或*Wan3.0*时，视频编辑器将完整的来源机器人视频片段作为运动参考，将编辑后的关键帧作为外观参考，并输入包含原子动作描述的文本提示。提供完整的机器人视频片段，可在整个生成过程中锚定动作时序、物体轨迹和状态转换，而无需显式使用分割掩码、深度图或边缘图控制。

**验证与配对**。视觉语言模型（VLM）验证器会筛选生成的视频片段，检查其视觉质量（完整手臂替换和场景稳定性）以及与来源轨迹的时间动作一致性；通过验证的视频片段会与原始机器人观测和动作配对（详情见附录E.4）。

## 4 方法

我们首先对人类视频条件下的上下文内操作进行形式化定义，随后介绍 Entangle 架构及其训练配方，该配方结合了任务内 InfoNCE 对比目标与可变长度窗口训练。

### 4.1 预备知识

**问题形式化**。我们考虑由单个人类演示视频$\mathcal{V}^{\text{ctx}} = (v_1^{\text{ctx}}, \dots, v_K^{\text{ctx}})$ 条件化的上下文内机器人操作，该视频指定了目标操作行为。给定$\mathcal{V}^{\text{ctx}}$和当前机器人视觉观测 $I_t$，策略通过建模它们的条件联合分布，预测未来机器人视频帧 $V = (I_{t+1}, \dots, I_{t+T_v})$ 以及可执行的动作片段 $A = (a_t, \dots, a_{t+T_a-1})$:

$$p_\theta\big(V, A \mid I_t, \mathcal{V}^{\text{ctx}}\big)\,. \tag{1}$$

联合建模未来视觉状态与动作，可在示范上下文下将视觉动态预测与控制结合起来。测试时，模型参数 $\theta$ 保持固定：机器人执行预测的动作前缀，并基于更新后的观测，使用同一示范$\mathcal{V}^{\text{ctx}}$进行闭环重新规划。

**条件流匹配**。 我们使用条件流匹配 [84]训练视频-动作骨干网络。对于每种模态 $m \in \{v, a\}$，将干净的未来视频潜变量 $X_0^v = \mathcal{E}(V)$ 或归一化动作 $X_0^a$与独立的高斯噪声 $\epsilon^m \sim \mathcal{N}(0, I)$进行插值，得到 $X_{\sigma_m}^m = (1-\sigma_m)X_0^m + \sigma_m \epsilon^m$, 其中 $\sigma_m \in [0, 1]$。在给定人类上下文和当前机器人观测的条件下，模型使用以下方式预测目标速度 $\epsilon^m - X_0^m$

$$\begin{aligned} \mathcal{L}_m &= \mathbb{E}\big[\|u_\theta^m - (\epsilon^m - X_0^m)\|_{M_m}^2\big]\,, \\ \mathcal{L}_{\text{WAM}} &= \mathcal{L}_v + \lambda_a \mathcal{L}_a, \end{aligned} \tag{2}$$

其中， $M_m$ 排除干净的条件标记， $\lambda_a$ 用于为动作损失加权。推理时，视频和动作样本从噪声中联合去噪，并解码为未来帧和可执行动作。

### 4.2 模型架构

图 3（a）展示了 Entangle 架构。人工演示 $\mathcal{V}^{\text{ctx}}$ 和机器人视频 （$I_t$、 $V$）由共享的 3D 变分自编码器$\mathcal{E}$ 编码为时空潜在网格，分别包含 $F_c$ 和 $F_q$ 帧，其中演示潜变量和当前观测 $I_t$ 保持干净（$\sigma_v = 0$），而未来机器人潜变量 $X_{\sigma_v}^v$则添加了噪声。分块视频令牌形成统一流 $H_V$，并与投影动作令牌 $H_A$一同处理。

**采用定向注意力的混合变换器** $V \to A$ 。遵循混合变换器（MoT）设计， [9, 10, 46], 我们的世界–动作扩散变换器骨干网络为视频流和动作流分别保留模态专属的 QKV 投影、归一化层和前馈网络，仅通过共享注意力将二者耦合起来。与基础模型一样，早期模块独立处理两条流；在其余模块中，动作查询关注视频键，而视频查询从不关注动作键（图 3（a））。这种不对称性

使动作流能够读取上下文条件视觉特征，同时避免嘈杂的动作令牌干扰视频生成。

时间 RoPE 对齐。视频令牌在时间、高度和宽度坐标 $(t, h, w)$ 上使用三维旋转位置嵌入（RoPE），并结合相加式相机视角嵌入；动作令牌使用一维时间 RoPE。图 3(b) 中的三种布局仅在每个片段重置或偏移的轴上有所不同：时间重置共享空间网格 $(h, w)$，但为两个片段分配相同的 $t$；空间偏移使 $(h, w)$ 彼此分离，却破坏了跨片段对应关系；我们采用的时间偏移使 $t$ 彼此分离，同时共享 $(h, w)$。这种布局沿时间轴依次排列人类上下文（$t \in \{0, \ldots, F_c - 1\}$）和机器人视频（$t \in \{F_c, \ldots, F_c + F_q - 1\}$），在保持空间对应关系的同时避免时间冲突。相应地，动作时间位置锚定于机器人视频片段的起始位置，而非 $t = 0$（图 3(b)）：

$$\pi_A(k) = F_c + \frac{k}{\rho}, \tag{3}$$

其中，$k$ 是动作步索引，$\rho$ 是动作到潜变量的时间比例。将 $\pi_A(k)$ 动态平移 $F_c$，可使动作令牌在不同上下文长度下始终与机器人视频片段保持时间对齐。

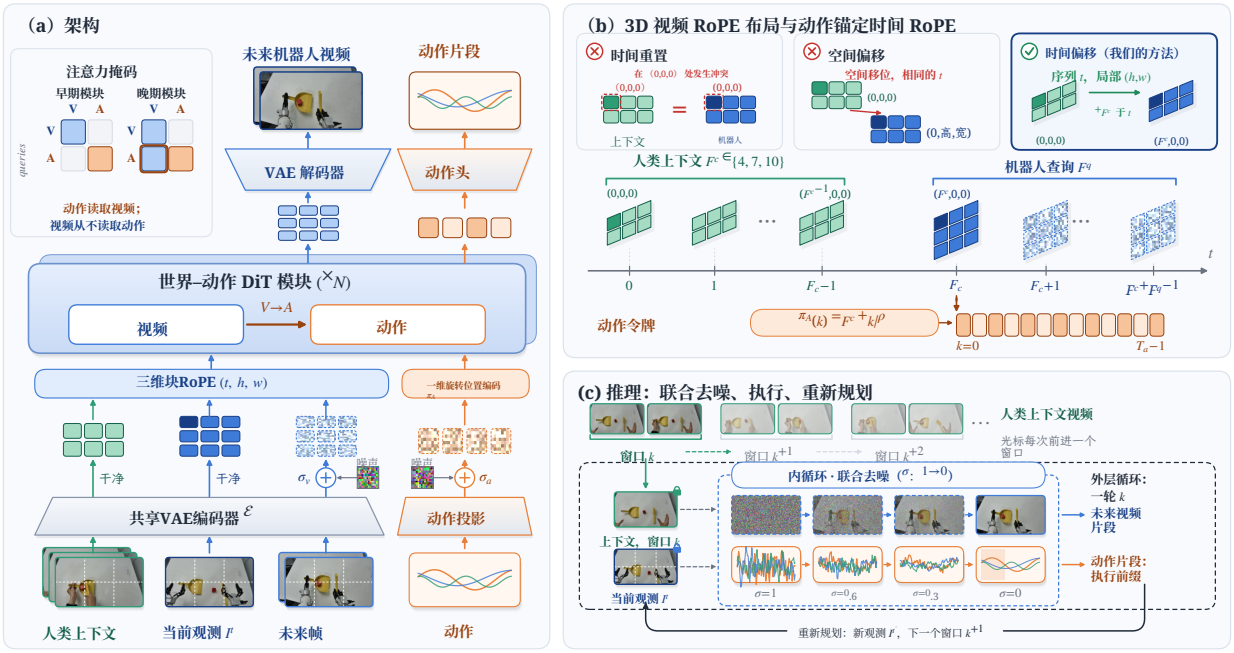



**图3 Entangle 概览**。（a）人类上下文和当前机器人观测共同构成视频—动作联合预测的条件，其中动作令牌关注视频令牌。（b）时间 RoPE 在保留局部空间坐标的同时，按顺序排列上下文和机器人视频，并将动作位置锚定到机器人片段。（c）每个推理轮次都会联合去噪未来视频和动作，执行动作前缀，并结合更新后的观测和上下文重新规划。

## 4.3 训练

我们从 MotuBrain [10] 开始，并在完整的 Ent-Dataset（第 3节）上继续训练（CT），采用以下目标。我们直接在真实机器人上评估 CT；RoboTwin 则使用多任务监督微调（SFT）进行模拟器适配。

**任务内 InfoNCE 对比目标**。流匹配损失 $\mathcal{L}_{\text{WAM}}$（公式2）对每条轨迹单独进行监督：当一个任务存在多种有效执行方式（例如不同的访问顺序或避障路径）时，它不会显式惩罚那些偏向同一任务其他有效轨迹的预测，因此策略可以依赖从 $I_t$ 推断出的粗略任务身份，而无需跟踪 $\mathcal{V}^{\text{ctx}}$。受负样本感知微调的启发 [14],，我们引入任务内 InfoNCE 目标 [13]，对比单步

$\widehat{}$干净动作估计 $A_i = X_i - \sigma_i u^a_{\theta,\ i}$（来自加噪动作输入 $X_i = X^a_{\sigma_i,\ i}$ 噪声水平为 $\sigma_i$）与批次中同一任务的其他$\widehat{}$轨迹。令 $\mathcal{G}_i$ 表示批次中的同任务样本（$|\mathcal{G}_i| \geq 2$），$d(A_i,\ A_j)$ $\widehat{}$ $\widehat{}$表示 $A_i$ 与干净轨迹 $A_j$之间的均方距离。我们将 $A_i$ 拉向其上下文配对轨迹$A_i$，同时排斥同任务的其他候选轨迹 $\{A_j\}_{j\in\mathcal{G}_i \setminus\{i\}}$，从而避免仅凭任务身份走捷径：

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} -\log \frac{\exp[-d(\widehat{A}_i, A_i)/\tau]}{\sum_{j\in\mathcal{G}_i}\exp[-d(\widehat{A}_i, A_j)/\tau]}, \tag{4}$$

其中，$\tau$ 为温度，$\mathcal{B}$ 为有效锚点集合，从而得到总损失 $\mathcal{L} = \mathcal{L}_{\text{WAM}} + \lambda_{\text{InfoNCE}}\mathcal{L}_{\text{InfoNCE}}$。

**可变长度窗口训练。** 上下文窗口时长决定了程序跟随与泛化之间的权衡：以完整回合片段为条件会促使模型推断全局任务身份，并从语义层面完成任务，而非跟随上下文；而仅使用短片段训练则会限制时间上下文，损害泛化能力。为平衡这一权衡，我们随机采样不同持续时间（13、25 或 37 帧，编码为 $F_c \in \{4,7,10\}$ 个潜在帧）的同步人类与机器人窗口，并配以匹配的动作片段；公式3 中的锚定式 RoPE 可确保动作在任意窗口长度下保持对齐。

**采用自适应时域（AH）的闭环推理。** 以开环方式执行预测动作片段时，随着未来不确定性增加，后续步骤可能逐渐失去与人类参考的关联，在下一次观测前不断累积接触和轨迹误差。因此，在每一轮重新规划中（图 3(c)），我们会监测 $V \to A$预测动作对参考标记的注意力：当后续步骤表现出较低的参考注意力权重和较高的时间熵（表明其与示范的关联微弱且分散）时，自适应时域（AH）会将执行截断至可靠前缀（6–24 步），按已执行的时长推进上下文窗口，并根据新观测重新规划。附录 E.3 详细介绍了 AH 和上下文同步。

| 组别 | 方法 | OS | SS | OM | MP | BB | EC | SB | 总体 |
|---|---|---|---|---|---|---|---|---|---|
| 基线 | WAN-Action [9] | $3.00 \pm 2.16$ | $7.33 \pm 1.25$ | $2.26 \pm 1.60$ | $10.67 \pm 1.70$ | $15.26 \pm 2.55$ | $38.33 \pm 2.05$ | $0.00 \pm 0.00$ | $10.98 \pm 1.07$ |
| | LingBot-VA [55] | $6.17 \pm 4.87$ | $3.67 \pm 2.49$ | $\underline{29.33} \pm 10.66$ | $23.33 \pm 8.22$ | $17.33 \pm 6.18$ | $42.33 \pm 7.85$ | $0.00 \pm 0.00$ | $17.45 \pm 1.40$ |
| | Zero-WAM [9] | $24.67 \pm 2.05$ | $47.00 \pm 4.55$ | $\mathbf{59.00} \pm 2.83$ | $69.14 \pm 2.93$ | $35.00 \pm 3.74$ | $\mathbf{84.87} \pm 0.18$ | $9.00 \pm 2.16$ | $46.95 \pm 0.72$ |
| | RoboICL* [85] | $\underline{53.33} \pm 25.17$ | $66.67 \pm 5.77$ | $3.33 \pm 5.77$ | $70.00 \pm 26.46$ | $20.00 \pm 10.00$ | $13.33 \pm 5.77$ | $0.00 \pm 0.00$ | $32.38 \pm 6.44$ |
| 消融实验 | 时间重置 | $\mathbf{62.00} \pm 2.65$ | $\underline{77.67} \pm 5.03$ | $3.00 \pm 1.73$ | $\underline{82.33} \pm 4.73$ | $\mathbf{94.00} \pm 3.00$ | $73.33 \pm 7.09$ | $\underline{61.00} \pm 1.73$ | $\underline{64.76} \pm 2.65$ |
| | 空间偏移 | $43.33 \pm 4.62$ | $71.33 \pm 2.31$ | $2.00 \pm 0.00$ | $75.67 \pm 3.06$ | $83.67 \pm 1.53$ | $65.33 \pm 1.15$ | $0.00 \pm 0.00$ | $48.76 \pm 1.25$ |
| | 不使用上下文学习 | $25.33 \pm 2.89$ | $56.33 \pm 8.08$ | $19.67 \pm 4.16$ | $57.67 \pm 3.06$ | $\underline{92.67} \pm 2.08$ | $69.33 \pm 4.51$ | $2.33 \pm 1.15$ | $46.19 \pm 1.92$ |
| *Ours* | Entangle（我们的方法） | | | | | | | | |

**表1** RoboTwin成功率（%，↑）：个随机种子的均值和 ± 标准差；最佳/次佳均值分别以粗体/下划线标示。任务：OS，物体放在秤上；SS，印章盖印；OM，打开微波炉；MP，订书机到垫板；BB，篮中面包；EC，杯垫上的空杯；SB，堆叠三个方块。WAN-Action、LingBot-VA和Zero-WAM的结果来自Zero-WAM表 2 [9]；*RoboICL报告首次尝试成功率（附录E.2）。

## 5 实验

我们在RoboTwin基准 [15, 16]以及实体Astribot S1机器人上，从两个泛化轴评估Entangle：**任务分布外**，即完整目标和多步骤流程在训练数据中均未出现；以及**上下文分布外**，即执行场景通过重新排列布局或替换物体和工具，偏离了严格对齐的训练对。

### 5.1 实验准备

**推理协议。** 仿真和真实机器人推理均仅以一个包含 13 帧的人类视频窗口为条件，不输入语言，使用 50 个去噪步数，视频/动作引导尺度分别为 5 和 1。**仿真（RoboTwin）。** 我们通过在 43 个 RoboTwin训练任务上进行联合监督微调来应对模拟器域偏移，然后在 7 个未见任务上进行评估。每轮预测后，上下文推进 13 帧。 **真实机器人**

（**Astribot S1**）。 我们直接评估 CT模型，不进行下游监督微调，以 10 Hz（2.4 秒）预测 24 个动作步，并通过自适应时域（AH）每轮执行 6–24 步（不使用 AH 时则执行全部 24 步；附录 E.3）。

**指标**。 我们报告仿真中的二元成功率，以及真实机器人上的分级任务得分（0–100）。真实机器人得分按每个任务和配置进行 10 次试验取平均，衡量是否同时完成演示目标和程序性约束（评分标准见附录A.3）。

## 5.2 RoboTwin基准评估

我们在 Zero-WAM `demo_clean` 协议 [9]下评估了七个未纳入监督微调训练的 RoboTwin任务：每个任务使用三个随机种子，每个随机种子运行 100 个回合，并为每个任务固定使用一个生成的人类上下文。Entangle 达到 69.52%，超过 Zero-WAM（46.95%）22.57 个百分点（表 1），并比基于智能体的 RoboICL（32.38%）高 37.14 个百分点。尽管 Zero-WAM 在单物体任务（例如打开微波炉、放置杯子）上表现良好，但其在三方块堆叠（SB）上的得分仅为 9.00%，而 Entangle 达到 70.33%。这一差距表明，由编辑后的初始帧生成的任务匹配片段足以支持短时、单阶段动作，而对于多步序列，帧对齐对应关系至关重要。

## 5.3 真实机器人任务分布外评估

这六项 Astribot S1 任务未出现在预训练和持续训练中。它们用于测试 **方块操作** (物体角色与顺序)、**鼠标使用** (新颖抓取、表面移动和点击)、**拧毛巾** (双手协调)、**反直觉目标** (非常规目的地)、**清扫** (使用杯子作为工具)，以及**避障** (沿指定通道行进，而非仅仅伸手到达目标；图 4)。附录 A 给出了任务定义，附录 C则将人类上下文与机器人执行配对。




**图 4** 六项真实机器人任务，执行帧按时间顺序排列。**(f)** 中的青色箭头标示了要求通过的左侧开口。

组件控制实验。 表 2 对严格配对、同任务轨迹对比（InfoNCE）、变长训练（VL）和自适应时域（AH）进行消融；完整模型在六项任务中的五项上表现最佳（均值为 83.5）。训练时将参考替换为另一个同任务回合（无配对，41.3），虽能在 Blocks 上传达粗略意图，却会导致精细双手扭转（Towel）、轨迹跟踪（Mouse）和反直觉放置上的表现崩溃：任务级配对无法监督精细动作或违背先验的目标。去掉 InfoNCE（64.7）后，策略会逐渐偏向一条与障碍物碰撞的平均路径，使多路径任务（Avoidance 和 Mouse）的表现下降。固定13帧训练（无 VL，62.7）将模型限制在短视局部映射中，无法将绕行通道与目标联系起来，导致避障得分降为零。最后，以开环方式执行全部 24 步（无 AH，61.3）会损害避障和拧毛巾任务，因为随着参考注意力减弱，继续执行后续步骤会累积路径和接触误差。因此，数据级对齐、训练级对比和部署级重新规划都不可或缺。

## 5.4 消融实验

**位置编码**。 在相同协议下，我们的时间偏移布局达到 69。52%，而时间重置为 64。76%，空间偏移为 48。76% （堆叠三个方块任务为零；表1）。20.76 个百分点的性能损失表明，在匹配视角下共享二维坐标十分重要：时间偏移将两个片段分开，同时保留了这种对应关系。

**表 2** Astribot S1 消融实验：10 次试验平均得分（0–100，↑）。每列的最佳/次佳结果：粗体/下划线。反直觉目标：反直觉目标；Avoid.：障碍避让。

| 配置 | 方块 | 鼠标 | Towel | 反直觉目标 | 清扫 | 避障 | Avg. |
|---|---|---|---|---|---|---|---|
| 无自适应时域 | <u>89</u> | 71 | 55 | 60 | 65 | 28 | 61.3 |
| 无 InfoNCE | <u>89</u> | 53 | **83** | <u>79</u> | 64 | 20 | <u>64.7</u> |
| 无 VL | 84 | <u>80</u> | 71 | 66 | <u>75</u> | 0 | 62.7 |
| 不进行配对 | 76 | 26 | 7 | 35 | 66 | <u>38</u> | 41.3 |
| Full | **92** | **96** | <u>76</u> | **91** | **84** | **62** | **83.5** |

**视频上下文的作用。** 一个单独训练、没有人类视频上下文的基线模型（无上下文学习）整体达到 46。19%；当目标可从初始场景中显而易见时，其表现出色（篮中面包任务达到 92。67%，高于完整模型的 90。33%），但在堆叠三个方块任务上降至 2。33%，而 Entangle 在该任务上达到 70。33%。这一模式表明，场景先验足以应对单阶段流程，却不足以完成需要从示范中推断的操作流程。

## 5.5 真实机器人任务分布外的数据扩展

我们在六个真实机器人任务上评估了五种嵌套的CT数据比例（0%、10%、25%、50%和100%）（见图 5），其中0%表示直接评估预训练骨干网络，非零比例则训练4,000个CT步数。在没有对齐的CT数据（0%）时，预训练骨干网络在所有任务上的得分均为零；随着对齐数据从10%扩展到100%，均值得分单调上升，从40.8升至83.5。在各项任务中，少量数据（10%–25%）已经足以让策略理解高层意图，例如物体排序（Blocks）、反直觉的目的地（Counterintuitive）和工具改用（Sweeping）；而持续的双手协调（Towel）和严格的路线遵循（Avoidance）则需要语料库中50%–100%的数据。这表明，尽管适量的配对数据能够教会策略 需要针对什么 物体或目标，而通过我们的合成流程扩展时间对齐数据

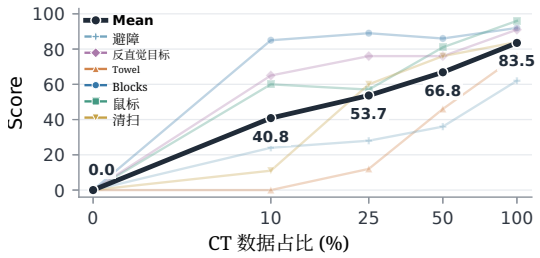



**图5** 六个真实机器人任务上的CT数据扩展（每项10次试验）。浅色曲线：任务得分；深色曲线：未加权平均值。正数据比例按对数间隔排列。

对于将人类如何移动转化为精细化机器人控制至关重要。

## 5.6 上下文分布外评估

由于所有训练对都严格进行场景对齐，我们评估了四种难度逐步增加的上下文分布外偏移（图6），以考察当工作空间偏离参考场景时，Entangle 如何建立对应关系。在**空间布局偏移**（重新定向的笔）和**类别内替换**（从黑色鼠标替换为白色鼠标）条件下，策略会调整其抓取方向，并在不同鼠标外观之间复现相同的描字轨迹。在**保持可供性不变的跨物体迁移**和**可供性重新利用**条件下，策略会跨物体类别迁移交互：将按住并擦拭的动作从相框迁移到支架上，并在缺少刷子和簸箕时，利用香蕉将木块清扫到盒子中。因此，在世界—动作先验上监督严格的时间对齐，并不会将策略锁定在像素级复制上；相反，Entangle 展现出一种涌现能力，能够将示范交互迁移到新颖的物体和场景中，使部署不再受严格场景对齐的限制，从而适应多样化的示范。

## 5.7 定性分析

### 5.7.1 使用和不使用 InfoNCE

**布料旋转与抛掷。** 两个策略都接收相同的参考，其中人类抬起一块布，做圆周运动，并将其朝相框抛去。使用 InfoNCE 时，手腕旋转更加匹配

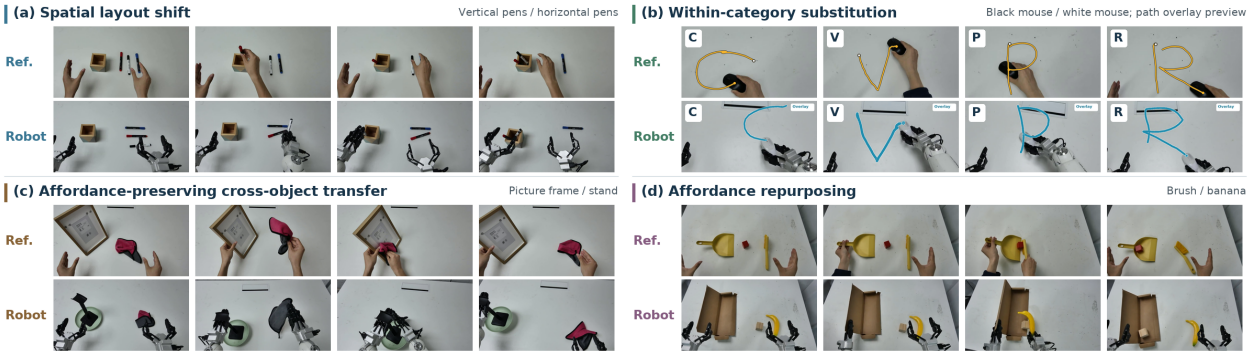



**图6** 四种上下文分布外情况，以紧凑的过程序列呈现。(a) 空间布局发生变化时的钢笔摆放；所选帧展示了初始场景和前两次摆放。(b) 黑色鼠标替换为白色鼠标：线条表示鼠标书写笔画。(c) 相框到支架的转移保留了按住并擦拭交互。(d) 香蕉承担了清扫工具这一非常规角色。

该示范。机器人张开夹爪，但未能抛出布。未使用 InfoNCE 时，机器人仅部分跟随示范中的圆周运动，但成功将布抛出（图7）。




**图7** 在同一人类视频条件下的两次布料旋转与抛掷运行轨迹。示意箭头展示转动阶段的圆周运动或左右运动，以及最后一列中的抛掷或掉落方向。两次运行轨迹中的布料颜色不同。

**擦拭相框背面。** 两个策略接收相同的参考视频，其中人类翻转相框，用左手将其稳住，同时用右手擦拭，然后收回双手。采用 InfoNCE 时，左夹爪在整个擦拭过程中都夹持着相框。随后，机器人按照上下文视频所示收回机械臂。在收回过程中，相框被其中一个夹爪钩住。未采用 InfoNCE 时，左夹爪过早收回，右夹爪在边缘附近擦拭，却没有演示中的支撑（图8）。

任务内 InfoNCE 会在同一任务的候选轨迹中，优先选择与当前视频配对的轨迹。在这些示例中，使用 InfoNCE 训练的策略更贴近演示的操作流程。未使用该方法的运行结果展现了采取替代轨迹时的灵活性与代价。不同的腕部运动成功完成了投掷，而擦拭过程中过早收回则省略了演示中的支撑动作。

### 5.7.2 使用生成的视频上下文进行部署

我们以 Seedance 生成的视频作为 Astribot S1 的条件输入。视频显示，左手将笔筒向前推，右手在旁边收拢三支笔。机器人以零样本方式成功执行了这一两步操作流程（图 9）。

此示例展示了视频生成模型如何充当零样本视觉规划器 [12, 82]。生成器会生成视觉计划，而策略则将该计划落实到机器人的工作空间中，从而实现任务




**图8** 圆圈标出了画面卡顿（中间一行）和左夹爪过早撤回（底下一行）。

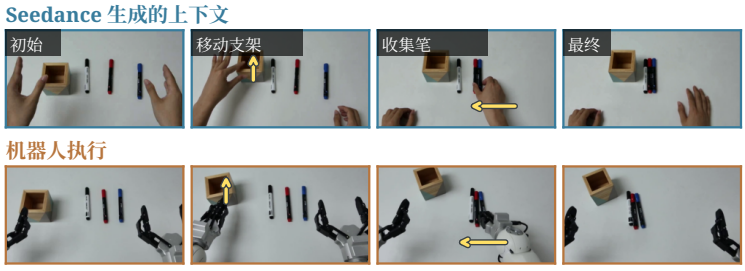



**图9** 在 Seedance 生成的上下文中部署，未录制人工演示。Astribot S1 先推笔筒，然后收拢三支笔。黄色箭头表示运动方向。

规范，无需录制人工演示。它可以让机器人自主执行策略训练期间未见过的各种任务，从而扩大真实机器人数据的收集规模。

### 5.7.3 无任务物体的执行

人类用右手拿起一支笔，递给左手，然后将笔放入托架。在空桌面上，机器人合上右夹爪，将两个夹爪移到一起，并张开右夹爪，同时保持左夹爪闭合。随后，它在放置动作结束时张开左夹爪，从而保留了示范中的动作顺序和手部角色（图 10）。




**图10** 机器人在空桌面上重现了示范的递交动作。

在笔和托架都不存在时，场景中没有可用于迁移的物体状态线索。保留下来的递交顺序和释放顺序展示了策略遵循上下文视频的能力。

## 6 局限性与未来工作

我们的真实机器人评估仅限于一种具身形态，而且数据流水线依赖于生成的人类视频的保真度。跨场景迁移和生成上下文部署通过定性案例进行了展示；仍需开展更广泛的定量评估。未来的工作将探索跨具身形态迁移、更可靠的数据生成，以及针对更长任务的故障恢复。

## 7 结论

我们介绍了 Entangle。该方法使用人类视频作为指定机器人行为的接口，无需针对任务进行参数更新。我们的研究结果表明，从对齐的人机经验中学习，可以支持超越熟悉任务和场景的可迁移示范跟随。这为实现更易于使用的机器人适应提供了一条途径，使人们能够通过展示行为应如何执行来传达新的行为。

## 参考

[1] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。载于神经信息处理系统进展，第 33 卷，第 1877–1901 页，2020年。[2] Vidhi Jain、Maria Attarian、Nikhil J Joshi、Ayzaan Wahid、Danny Driess、Quan Vuong、Pannag R Sanketi、Pierre Sermanet、Stefan Welker、Christine Chan 等。Vid2Robot：利用交叉注意力变换器进行端到端视频条件策略学习。载于机器人学：科学与系统，2024年。[3] Max Fu、Huang Huang、Gaurav Datta、Lawrence Yunliang Chen、Will Panitch、Fangchen Liu、Hui Li 和 Ken Goldberg。ICRT：通过下一词元预测实现上下文内模仿学习。载于*2025*年 *IEEE*机器人与自动化国际会议（*ICRA*），第 5937–5944 页。IEEE，2025年。[4] Vitalis Vosylius 和 Edward Johns。Instant Policy：通过图扩散实现上下文内模仿学习。载于 Y. Yue、A. Garg、N. Peng、F. Sha 和 R. Yu（编），国际学习表征会议，第 2025 卷，第 35011–35030 页，2025年。[5] Kaustubh Sridhar、Souradeep Dutta、Dinesh Jayaraman 和 Insup Lee。RICL：为预训练视觉-语言-动作模型添加上下文内适应能力。载于机器人学习会议，第 305 卷，机器学习研究论文集，第 5022–5038 页，2025年。[6] Chengwen Zhang、Jianyang Wang、Chang Liu、Yukun Wang、Yuan Xu、Zhoutong Ye、Jie Cai、Chun Yu 和 Yuanchun Shi。通过联合行动理解人机协作。*arXiv* 预印本 *arXiv:2610.37298*，2026年。[7] Seonghyeon Ye、Yunhao Ge、Kaiyuan Zheng、Shenyuan Gao、Sihyun Yu、George Kurian、Suneel Indupuru、You Liang Tan、Chuning Zhu、Jiannan Xiang 等。世界动作模型是零样本策略。*arXiv* 预印本 *arXiv:2602.15922*，2026年。[8] Qihang Zhang、Lin Li、Luyao Zhang、Shuai Yang、Yiming Luo、Shuaiting Li、Ruilin Wang、Junke Wang、Jiahao Shao、Gangwei Xu 等。面向可泛化机器人控制的原生视频-动作预训练。*arXiv* 预印本 *arXiv:2607.08639*，2026年。[9] Jiaming Zhou、Qihang Zhang、Gangwei Xu、Cunxin Fan、Yujie Zhao、Ruilin Wang、Yiming Luo、Shuai Yang、Xing Zhu、Yujun Shen、Junwei Liang 和 Yinghao Xu。Zero-WAM：基于人类视频的上下文内世界动作建模，实现开放式任务泛化。*arXiv* 预印本 *arXiv:2608.26103*，2026年。[10] Motubrain Team、Chendong Xiang、Fan Bao、Haitian Liu、Hengkai Tan、Hongzhe Bi、James Li、Jiabao Liu、Jingrui Pang、Kiro Jing、Louis Liu、Mengchen Cai、Rongxu Cui、Ruowen Zhao、Runqing Wang、Shuhe Huang、Yao Feng、Yinze Rong、Zeyuan Wang 和 Jun Zhu。Motubrain：用于机器人控制的先进世界动作模型。*arXiv* 预印本 *arXiv:2604.27792*，2026年。[11] Yao Feng、Chendong Xiang、Xinyi Mao、Hengkai Tan、Zuyue Zhang、Shuhe Huang、Kaiwen Zheng、Haitian Liu、Hang Su 和 Jun Zhu。Vidarc：用于闭环控制的具身视频扩散模型。*arXiv* 预印本 *arXiv:2512.17661*，2025年。

[12] Fan Bao、Chendong Xiang、Gang Yue、Guande He、Hongzhou Zhu、Kaiwen Zheng、Min Zhao、Shilong Liu、Yaole Wang 和 Jun Zhu。Vidu：一种基于扩散模型、具备高度一致性、动态性和熟练度的文本到视频生成器。*arXiv* 预印本 *arXiv:2405.04233*，2024年。[13] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。利用对比预测编码进行表征学习。*arXiv* 预印本 *arXiv:1807.03748*，2018年。[14] Kaiwen Zheng、Huayu Chen、Haotian Ye、Haoxiang Wang、Qinsheng Zhang、Kai Jiang、Hang Su、Stefano Ermon、Jun Zhu 和 Ming-Yu Liu。DiffusionNFT：利用前向过程进行在线扩散强化学习。载于国际学习表征会议，2026年。[15] Yao Mu、Tianxing Chen、Zanxin Chen、Shijia Peng、Zhiqian Lan、Zeyu Gao、Zhixuan Liang、Qiaojun Yu、Yude Zou、Mingkun Xu、Lunkai Lin、Zhiqiang Xie、Mingyu Ding 和 Ping Luo。RoboTwin：基于生成式数字孪生的双臂机器人基准。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 27649–27660 页，2025年。[16] Tianxing Chen、Zanxin Chen、Baijun Chen、Zijian Cai、Yibin Liu、Zixuan Li、Qiwei Liang、Xianliang Lin、Yiheng Ge、Zhenyu Gu 等。RoboTwin 2.0：具备强领域随机化的可扩展数据生成器与基准，适用于鲁棒双手机器人操作。*arXiv* 预印本 *arXiv:2506.18088*，2025年。[17] Yan Duan、Marcin Andrychowicz、Bradly Stadie、Jonathan Ho、Jonas Schneider、Ilya Sutskever、Pieter Abbeel 和 Wojciech Zaremba。单样本模仿学习。载于神经信息处理系统进展，第 30 卷，2017年。[18] Stephen James、Michael Bloesch 和 Andrew J. Davison。用于少样本模仿学习的任务嵌入式控制网络。载于第二届机器人学习会议论文集，第 783–795 页，2018年。[19] Sudeep Dasari 和 Abhinav Gupta。用于单次视觉模仿的 Transformer。载于*2020*年机器人学习会议论文集，第 2071–2084 页，2021年。[20] Zhao Mandi、Fangchen Liu、Kimin Lee 和 Pieter Abbeel。迈向更具泛化能力的单次视觉模仿学习。载于*IEEE*机器人与自动化国际会议，第 2434–2444 页，2022年。数字对象标识符：10.1109/ICRA 46639.2022.9812450。[21] Tianhe Yu、Pieter Abbeel、Sergey Levine 和 Chelsea Finn。通过示范单样本组合基于视觉的技能。载于*IEEE/RSJ*智能机器人与系统国际会议，第 2643–2650 页，2019年。数字对象标识符：10.1109/IROS40897.2019.8967745。[22] Mengda Xu、Zhenjia Xu、Cheng Chi、Manuela Veloso 和 Shuran Song。Xskill：跨本体技能发现。载于机器人学习会议，第 3536–3555 页。PMLR，2023年。[23] Xiang Zhu、Yichen Liu、Hezhong Li 和 Jianyu Chen。将人类示范视频作为提示来学习可泛化的机器人策略。载于*IEEE*机器人与自动化国际会议（*ICRA*），2026年。[24] Suraj Nair、Aravind Rajeswaran、Vikash Kumar、Chelsea Finn 和 Abhinav Gupta。R3M：用于机器人操作的通用视觉表征。载于第六届机器人学习会议论文集，第 205 卷，机器学习研究论文集，第 892–909 页，2023年。[25] Seonghyeon Ye、Joel Jang、Byeongguk Jeon、Se June Joo、Jian wei Yang、Baolin Peng、Ajay Mandlekar、Reuben Tan、Yu-Wei Chao、Bill Yuchen Lin、Lars Liden、Kimin Lee、Jianfeng Gao、Luke Zettlemoyer、Dieter Fox 和 Minjoon Seo。基于视频的潜在动作预训练。载于国际学习表征会议，2025年。[26] Chen Wang、Linxi Fan、Jiankai Sun、Ruohan Zhang、Li Fei-Fei、Danfei Xu、Yuke Zhu 和 Anima Anandkumar。MimicPlay：通过观察人类操作实现长时程模仿学习。载于第七届机器人学习会议论文集，第 229 卷，机器学习研究论文集，第 201–221 页，2023年。[27] Simar Kareer、Dhruv Patel、Ryan Punamiya、Pranay Mathur、Shuo Cheng、Chen Wang、Judy Hoffman 和 Danfei Xu。EgoMimic：利用第一人称视角视频扩展模仿学习。载于*IEEE*机器人与自动化国际会议（*ICRA*），第 13226–13233 页，2025年。数字对象标识符：10.1109/ICRA55743.2025.11127989。[28] Kushal Kedia、Prithwish Dan、Angela Chao、Maximus A Pace 和 Sanjiban Choudhury。执行不匹配情况下的单样本模仿。载于*2025*年 *IEEE*机器人与自动化国际会议（*ICRA*），第 15649–15656 页。IEEE，2025年。

[29] Rutav Shah, Shuijing Liu, Qi Wang, Zhenyu Jiang, Sateesh Kumar, Mingyo Seo, Roberto Martín-Martín, and Yuke Zhu. MimicDroid：基于人类操作视频的人形机器人操作上下文学习。发表于*IEEE*机器人与自动化国际会议（*ICRA*），第15916–15923页，2026年。[30] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. 为未见操作任务学习视频条件策略。发表于*IEEE*机器人与自动化国际会议，2023年。[31] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z：通过机器人模仿学习实现零样本任务泛化。发表于第*5*届机器人学习会议论文集，第164卷机器学习研究论文集，第991–1002页。PMLR，2022年。[32] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 上下文学习在上下文中“学到”了什么：解耦任务识别与任务学习。发表于计算语言学协会：*ACL 2023*论文集，第8298–8319页，2023年。数字对象标识符：10.18653/v1/2023.findings-acl.527。[33] Raktim Goswami, Prashanth Krishnamurthy, Yann LeCun, and Farshad Khorrami. Osvi-wm：通过世界模型引导的轨迹生成，对未见任务进行单次视觉模仿。发表于神经信息处理系统进展，第38卷：54725–54745，2025年。[34] Cheng Qian, Ruomeng Fan, Yifei Ren, Yilong Wang, and Edward Johns. Synthicl：利用合成数据实现可扩展的上下文内模仿学习。发表于*arXiv*预印本 *arXiv:2606.08154*，2026年。[35] Jaehyeon Son, Junhyun Kim, Kyle Kam, Jeremiah Coholich, Seok Joon Kim, Jinhoo Kim, Chris Dongjoo Kim, Jaemin Cho, Dieter Fox, and Zsolt Kira. Seetraceact：基于跨具身演示视频的可见性感知潜在规划。发表于*arXiv*预印本 *arXiv:2606.02745*，2026年。[36] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. 时间对比网络：基于视频的自监督学习。发表于*IEEE*机器人与自动化国际会议（*ICRA*），第1134–1141页，2018年。数字对象标识符：10.1109/ICRA.2018.8462891。[37] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 时间循环一致性学习。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第1801–1810页，2019年。数字对象标识符：10.1109/CVPR.2019.00190。[38] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. XIRL：跨具身逆强化学习。发表于第*5*届机器人学习会议论文集，第164卷机器学习研究论文集，第537–546页。PMLR，2022年。[39] Guangyan Chen, Meiling Wang, Te Cui, Zichen Zhou, Qi Shao, Shalfun Li, Hang Su, Roy Gan, Hao Wang, Mengyin Fu, Yi Yang, and Yufeng Yue. HOST：机器人通过单个人类视频在数秒内习得操作技能，2026年。[40] Jiazhao Liang, Zhewen He, Shuaihang Yuan, Hao Huang, and Yi Fang. JEDI：用于上下文内模仿学习的联合嵌入演示压缩。发表于*arXiv*预印本 *arXiv:2610.40356*，2026年。[41] Minxing Li, Minghao Han, Weizhi Zhao, Hanwen Wang, Xiangshuo Liu, Shuyao Shang, Jingxiang Zhou, Mingchao Sun, Hongyu Pan, Mu Xu, Yu Liu, Lue Fan, and Zhaoxiang Zhang. 简化上下文内机器人学习：操作任务的普及化方案。发表于*arXiv*预印本 *arXiv:2609.38173*，2026年。[42] Xunzhe Zhou, Yiyang Cai, Fengyi Wang, Ran Ju, Hanxiang Ren, Ruizhe Liu, Yu Zhang, Qian Luo, Feng Chen, Pei Zhou, Yi Ma, and Yanchao Yang. The Imitator Game：超越动作预测的机器人模仿能力基准测试。发表于机器人学习会议，2026年。[43] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. 释放大规模视频生成式预训练在视觉机器人操作中的潜力。发表于国际学习表征会议，2024年。[44] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. 统一视频动作模型。发表于*arXiv*预印本 *arXiv:2503.00200*，2025年。[45] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. 统一世界模型：结合视频与动作扩散，在大型机器人数据集上进行预训练。发表于*arXiv*预印本 *arXiv:2504.02792*，2025年。

[46] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong等。Motus：统一潜在动作世界模型。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第35101–35113页，2026年。[47] Chelsea Finn and Sergey Levine. 用于机器人运动规划的深度视觉前瞻。发表于*IEEE*机器人与自动化国际会议，2017年。[48] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 通过文本引导的视频生成学习通用策略。发表于神经信息处理系统进展，第36卷：9156–9172，2023年。[49] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. 通过稠密对应关系从无动作视频中学习行动。发表于国际学习表征会议，2024年。[50] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate：通过视频生成实现真实世界视觉运动策略学习。发表于*arXiv*预印本 *arXiv:2406.16862*，2024年。[51] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. 预测性逆动力学模型是可扩展的机器人操作学习器。发表于国际学习表征会议，2025年。[52] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. 视频预测策略：具有预测性视觉表征的通用机器人策略。发表于国际机器学习会议，第24328–24346页，2025年。[53] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE：通过隐式世界建模进行机器人学习。发表于机器人学习会议，第3952–3971页，2025年。[54] Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. Videovla：视频生成器能够成为可泛化的机器人操作器。发表于神经信息处理系统进展，第38卷：95597–95621，2025年。[55] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. 面向机器人控制的因果世界建模。发表于*arXiv*预印本 *arXiv:2601.21998*，2026年。[56] Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar：用于通用操作的具身视频扩散模型。发表于*arXiv*预印本 *arXiv:2507.12898*，2025年。[57] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen：通过视频世界模型释放机器人学习中的泛化能力。发表于机器人学习会议，第5170–5194页，2025年。[58] Siyin Wang, Junhao Shi, Senyu Fei, Zhaoyang Fu, Li Ji, Jingjing Gong, and Xipeng Qiu. 面向机器人控制的上下文内世界建模。发表于*arXiv*预印本 *arXiv:2606.26025*，2026年。[59] Jeongeun Park, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. 从任务配对的人类视频中学习可迁移的世界-动作模型。发表于*arXiv*预印本 *arXiv:2610.14351*，2026年。[60] Haisheng Su, Zongdai Liu, Xin Jin, Haoxuan Dou, Chengming Hu, Baorun Li, Zhanwang Liu, Ruiyan Xu, Jianjie Fang, Xin Zhang, Zhenjie Yang, Xue Yang, Chen Gao, Junchi Yan, Yong Li, and Wei Wu. WorldScape Policy 2.0：利用推理增强记忆与上下文学习，赋能可控世界动作建模。发表于*arXiv*预印本 *arXiv:2607.18840*，2026年。[61] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. 通过元学习实现单次视觉模仿学习。发表于第*1*届机器人学习年度会议论文集，第357–368页，2017年。[62] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. 通过领域自适应元学习，从观察人类行为中进行单样本模仿。发表于*arXiv*预印本 *arXiv:1802.01557*，2018年。[63] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 通过自监督测试时训练，实现分布偏移下的泛化。发表于国际机器学习会议，第9229–9248页。PMLR，2020年。

[64] Nicklas Hansen、Rishabh Jangir、Yu Sun、Guillem Alenyà、Pieter Abbeel、Alexei A. Efros、Lerrel Pinto 和 Xiaolong Wang。部署期间的自监督策略适应。在国际学习表征会议，2021年。[65] Yu Sun、Xinhao Li、Karan Dalal、Jiarui Xu、Arjun Vikram、Genghan Zhang、Yann Dubois、Xinlei Chen、Xiaolong Wang、Sanmi Koyejo、Tatsunori Hashimoto 和 Carlos Guestrin。在测试时学习（学习如何在测试时学习）：具有表达性隐藏状态的循环神经网络。在 国际机器学习会议，第267卷机器学习研究论文集，第57503–57522页，2025年。[66] Yusen Feng、Bingchen Han、Jiangran Lyu、Kai Liu、Yixin Zheng、Yuxuan Wan、Weiheng Liu、Sun Han、Ruiqin Li、Yulong Zhang 等。Wam-ttt：通过在测试时观看人类操作来引导世界-动作模型。*arXiv*预印本 *arXiv:2607.06988*，2026年。[67] Yunfan Jiang、Yevgen Chebotar、Ruijie Zheng、Fengyuan Hu、Yunhao Ge、Jimmy Wu、Tianyuan Dai、Scott Reed、Li Fei-Fei、Yuke Zhu 等。Robottt：机器人策略的上下文扩展。*arXiv*预印本 *arXiv:2607.15275*，2026年。[68] Nick Heppert、Max Argus、Tim Welschehold、Thomas Brox 和 Abhinav Valada。DITTO：通过轨迹变换进行示范模仿。在 *IEEE/RSJ*智能机器人与系统国际会议（*IROS*），2024年。[69] Yifeng Zhu、Arisrei Lim、Peter Stone 和 Yuke Zhu。利用单个人类视频和开放世界物体图进行基于视觉的操作。*arXiv*预印本 *arXiv:2405.20321*，2024年。[70] Jinhan Li、Yifeng Zhu、Yuqi Xie、Zhenyu Jiang、Mingyo Seo、Georgios Pavlakos 和 Yuke Zhu。OKAMI：通过单视频模仿教授人形机器人操作技能。在 机器人学习会议，2024年。[71] Sungjae Park、Homanga Bharadhwaj 和 Shubham Tulsiani。DemoDiffusion：使用预训练扩散策略进行单次人类模仿。在 *IEEE*机器人与自动化国际会议（*ICRA*），2026年。[72] Yuzhe Qin、Yueh-Hua Wu、Shaowei Liu、Hanwen Jiang、Ruihan Yang、Yang Fu 和 Xiaolong Wang。DexMV：从人类视频中进行灵巧操作的模仿学习。在 欧洲计算机视觉会议，2022年。[73] Bharat Lal Bhatnagar、Xianghui Xie、Ilya A. Petrov、Cristian Sminchisescu、Christian Theobalt 和 Gerard Pons-Moll。BEHAVE：用于跟踪人类物体交互的数据集与方法。在 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第15935–15946页，2022年。数字对象标识符：10.1109/CVPR52688.2022.01547。[74] Yun Liu、Chengwen Zhang、Ruofan Xing、Bingda Tang、Bowen Yang 和 Li Yi。CORE4D：用于协作物体重新排列的四维人-物-人交互数据集。在 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第1769–1782页，2025年。数字对象标识符：10.1109/CVPR52734.2025.00172。[75] Pratyusha Sharma、Lekha Mohan、Lerrel Pinto 和 Abhinav Gupta。多重交互变得简单（MIME）：用于模仿的大规模示范数据。在 第*2*届机器人学习会议论文集，第87卷机器学习研究论文集，第906–915页。PMLR，2018年。[76] Hao-Shu Fang、Hongjie Fang、Zhenyu Tang、Jirong Liu、Chenxi Wang、Junbo Wang、Haoyi Zhu 和 Cewu Lu。RH20T：用于一次性学习多样技能的综合机器人数据集。在*IEEE*机器人与自动化国际会议（*ICRA*），第653–660页，2024年。数字对象标识符：10.1109/ICRA57147.2024.10611615。[77] Sicheng Xie、Haidong Cao、Zejia Weng、Zhen Xing、Haoran Chen、Shiwei Shen、Jiaqi Leng、Zuxuan Wu 和 Yu-Gang Jiang。Human2Robot：从配对人-机器人视频中学习机器人动作。*arXiv*预印本 *arXiv:2502.16587*，2025年。[78] Marion Lepert、Jiaying Fang 和 Jeannette Bohg。Phantom：仅使用人类视频、无需机器人即可训练机器人。在 机器人学习会议，第305卷机器学习研究论文集，第4545–4565页，2025年。[79] Yangcen Liu、Shuo Cheng、Xinchen Yin、Woo Chul Shin、Alfred Cueva、Yiran Yang、Zhenyang Chen、Chuye Zhang 和 Danfei Xu。EgoEngine：从第一人称人类视频到高保真灵巧机器人示范。*arXiv*预印本 *arXiv:2606.12604*，2026年。[80] Seungyong Lee、Suhyeok Jang、Minyoung Kim 和 Jinwoo Shin。通过虚拟试穿从人类视频合成灵巧机器人示范。*arXiv*预印本 *arXiv:2611.05163*，2026年。[81] Guangyan Chen、Meiling Wang、Qi Shao、Zichen Zhou、Weixin Mao、Te Cui、Minzhao Zhu、Yinan Deng、Luojie Yang、Zhanqi Zhang、Yi Yang、Hua Chen 和 Yufeng Yue。看一次，然后行动：从单次视频示范中进行任务学习的视觉-语言-动作模型。*arXiv*预印本 *arXiv:2512.07582*，2025年。

[82] Homanga Bharadhwaj、Debidatta Dwibedi、Abhinav Gupta、Shubham Tulsiani、Carl Doersch、Ted Xiao、Dhruv Shah、Fei Xia、Dorsa Sadigh 和 Sean Kirmani。Gen2Act：新颖场景中的人类视频生成能够实现可泛化的机器人操作。在 第*9*届机器人学习会议论文集，第305卷机器学习研究论文集，第3936–3951页。PMLR，2025年。[83] Alexander Kapitanov、Karina Kvanchiani、Alexander Nagaev、Roman Kraynov 和 Andrei Makhliarchuk。HaGRID——手势识别图像数据集。在 *IEEE/CVF*冬季计算机视觉应用会议论文集，第4572–4581页，2024年。[84] Yaron Lipman、Ricky T. Q. Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成式建模的流匹配。在国际学习表征会议，2023年。[85] Fangcheng Liu、Yeqing Shen、Anda Cheng、Weishi Mi、Chao Tang、Chenyuan Liu、Yushun Xiang、Tingguang Li、Yong-Lu Li 和 Yehui Tang。RoboICL：借助GPT-6 Astra进行具身上下文学习。*arXiv*预印本 *arXiv:2609.34261*，2026年。

# 附录

## A 真实机器人任务、环境与评分

### A.1 能力覆盖与任务选择

我们检查了 MotuBrain 的预训练和持续训练（CT）数据，核对每项评估任务的目标和完整流程。六项完整任务均未出现在训练数据集中。

表 3 指出了促使我们设置各项任务的主要能力。六项任务都要求进行多种操作，并可能涉及多项能力。

**表 3** 六项真实机器人任务所侧重的能力。

| 能力 | Task | 示范所指定的内容 |
|---|---|---|
| 识别与选择 | 方块操作 | 选择空间物体角色，并遵循左／右／中间的放置顺序。 |
| 新的物体交互 | 鼠标使用 | 按照演示进行抓取、表面移动和按钮按压，与鼠标进行交互。 |
| 动作细节获取 | 拧毛巾 | 复现演示中的折叠动作以及双手协同扭转动作。 |
| 反直觉目标 | 物体重新排列 | 遵循异常目的地和操作顺序。 |
| 新物体功能 | 杯子清扫 | 将杯子用作扫具，并用它把方块扫进簸箕。 |
| 流程约束 | 避障 | 使用指定的开口，并避免接触障碍物，同时获取并放置杯子。 |

### A.2 任务定义

**方块操作**。 将三个立方体堆叠在桌子上：先将初始位于左侧的立方体放在底部，再将初始位于右侧的立方体放在其上，最后将初始位于中间的立方体放在最上方。

**鼠标使用**。 抓取鼠标，在桌面上描画 C、V、P 和 R，按左键五次，然后将鼠标放回其起始区域。

**拧毛巾**。 抬起并折叠毛巾，用两个夹爪夹住相对的部分，向两个方向扭转，然后将毛巾放回桌子上。

**反直觉目标**。 按以下顺序完成操作：将方块从簸箕中倒出，将笔从笔筒移入簸箕，并将刷子放在笔筒顶部。然后将机械臂返回静止构型。不同试验之间物体的位置可能会略有变化。

**杯子清扫**。 抓取杯子和平底锅，用杯子将方块清扫进簸箕，抬起簸箕，然后将两种工具放回各自的起始区域。

**避障**。 通过左侧开口到达目标杯子，抓取它，并在避免接触障碍物的情况下将其稳定放置。障碍物的身份和位置可能因试验而异。

### A.3 特定任务评分标准

每次试验从100分开始，按照表4中的扣分项进行扣分，并将得分下限设为零。标有“每项”的扣分适用于表中指定的每个项目或事件（例如，每缺失或无法识别的字母扣10分）。其他所有扣分在每次试验中最多适用一次。如果机器人在任务上没有取得任何进展，则该次试验得零分。

**表4** 真实机器人评估的任务专属扣分规则。

| Task | 何时扣分 | 扣分<br>扣分 |
|---|---|---|
| **方块操作** | 规定堆叠中每缺少一个立方体 | 10 |
| | 左—右—中选择顺序错误，每次计一次 | 10 |
| **鼠标使用** | 每个缺失或无法识别的字母（C、V、P 或 R） | 10 |
| | 鼠标被抬起超过 5 厘米期间的每个字母 | 5 |
| | 鼠标被放下并重新抓取期间的每个字母 | 5 |
| | 五次按压中每漏按一次 | 10 |
| | 未按中左键的按压 , 计为一次 | 5 |
| | 未能将鼠标移回 | 5 |
| **拧毛巾** | 漏折一次 | 10 |
| | 每次失去抓取 | 5 |
| | 每次遗漏扭转方向 | 10 |
| | 最终放置位置不当 | 5 |
| **反直觉目标** | 每次重新摆放失败（倒出方块、将笔放入平底锅，或将刷子放在托架上） | 25 |
| | 未能将机械臂返回静止位置 | 20 |
| | 每次重新摆放出现轻微放置错误 | 10 |
| **杯子清扫** | 未能固定工具，计一次 | 10 |
| | 方块未转移到簸箕中，计一次 | 10 |
| | 未执行簸箕向上移动 | 10 |
| | 每件未归还至起始区域的工具 | 10 |
| **避障** | 每次接触到的不同障碍物 | 20 |
| | 未能抓取目标 | 20 |
| | 未能稳定放置目标 | 20 |
| | 未按规定从左侧通道通过的路线 | 30 |

## B 其他实验控制

**真实机器人**。自适应时域消融实验（**w/o AH**）使用相同的持续训练检查点，并在重新规划前执行全部24个预测动作（第E.3节）。对于InfoNCE消融实验（**w/o InfoNCE**），我们在训练期间省略对比损失项。固定长度训练对照（**w/o VL**）使用13帧窗口，替代由13帧、25帧和37帧窗口组成的混合设置。为消融严格配对（**w/o pairing**），我们将匹配的参考替换为来自同一任务不同回合的片段，同时保留原始目标轨迹和动作监督。我们在训练前固定替换分配。对于数据扩展，0%基线直接使用MotuBrain。10%、25%、50%和100%的训练集彼此嵌套，并共享相同的MotuBrain初始化、全局批次大小1,024、优化器设置以及4,000步持续训练预算。

**RoboTwin**。每种 RoPE 变体在训练和推理期间均采用其指定的位置布局。无参考模型从第 4,000 步的参考条件 CT 检查点开始，在没有人类视频上下文的情况下进行 4,000 步 RoboTwin 监督微调。评估时我们也省略参考。表5报告了三个评估随机种子的结果。

**表 5** RoboTwin 在三个评估随机种子下的控制结果。每个随机种子运行 700 个回合，每行使用一个固定检查点。我们报告成功次数，以及种子级成功率的均值 ± 样本标准差。

| 控制 | 种子 0 | 种子 1 | 种子 2 | 比率（%） |
|---|---|---|---|---|
| 时间偏移 | 491 | 487 | 482 | $69.52 \pm 0.64$ |
| 时间重置 | 432 | 462 | 466 | $64.76 \pm 2.65$ |
| 空间偏移 | 351 | 334 | 339 | $48.76 \pm 1.25$ |
| 无参考 | 308 | 333 | 329 | $46.19 \pm 1.92$ |

## C 真实机器人环境与执行可视化

图 11 将六项任务中每项的人类演示与机器人头部相机拍摄的执行帧配对。机器人头部相机和腕部相机的录制分辨率分别为 $1280 \times 720$ 和 $640 \times 360$，帧率为 10 fps。我们使用第 A.3 节中的评分标准对完整回合进行评分。

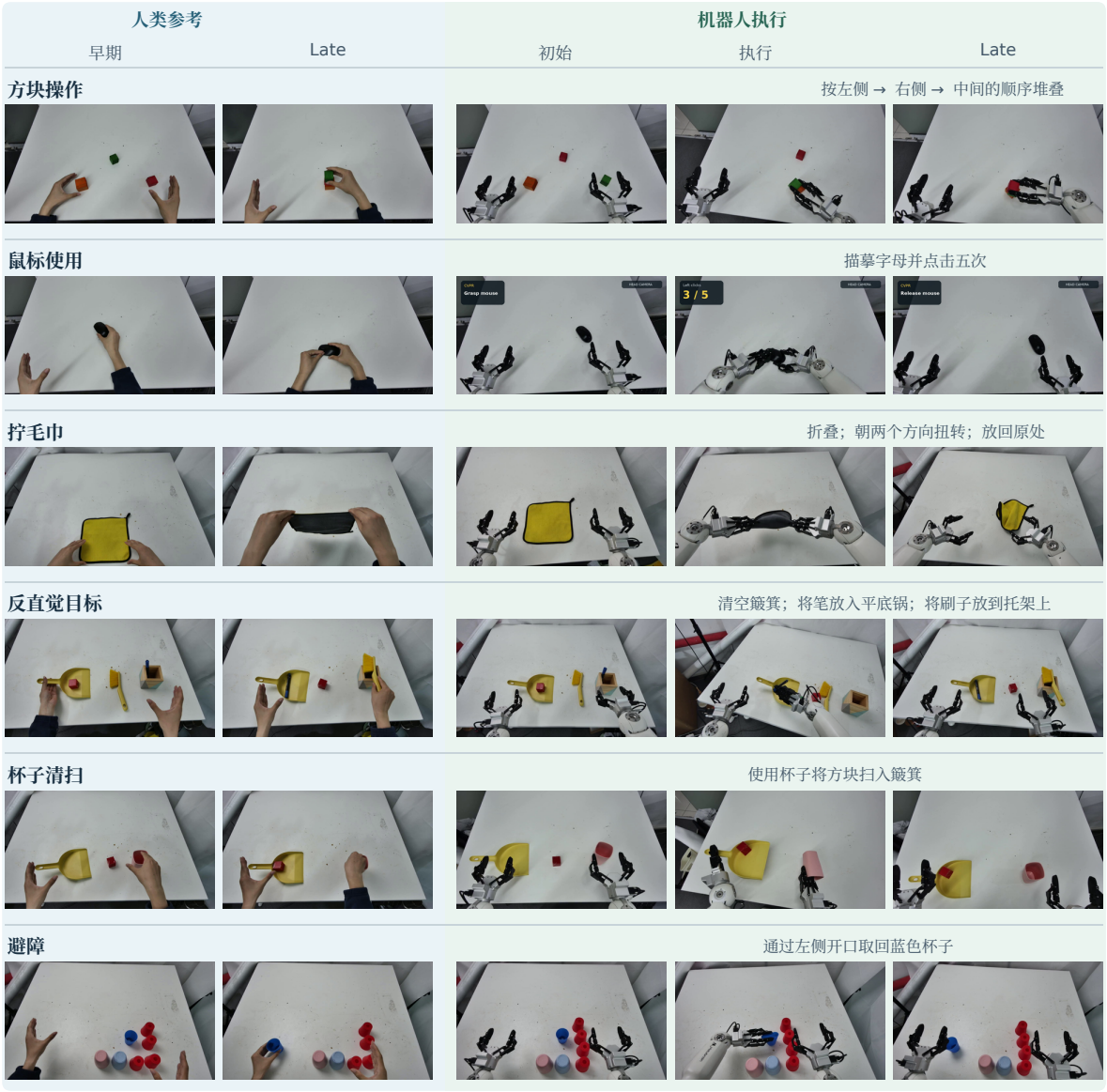



**图11** 六项真实机器人任务的人类演示和机器人执行。每行展示早期和晚期的演示帧，随后是机器人头部相机拍摄的初始、中间和晚期帧。

## D RoboTwin任务、环境与评分

### D.1 任务环境、目标与评估协议

我们遵循已发布的 Zero-WAM `demo_clean` 协议，在从43项任务仿真SFT划分中留出的七个 RoboTwin任务上进行实验。每项任务均使用一个固定生成的人类视频参考。对于每个视频条件检查点，我们针对每个任务、每个随机种子评估100个专家可完成场景。如果脚本专家

能够规划并完成该任务，则该场景符合要求。我们报告三个评估随机种子上的均值和样本标准差，并在每个随机种子内对七项任务的比率取平均，以得到总体得分。

**任务环境和目标。** 任务定义和场景准备与原始RoboTwin基准完全相同。

**OS：物体放在秤上**（将_物体_秤上）。鼠标、订书机或铃铛与电子秤放置在桌子的同一侧。物体的初始中心位于 $x \in [-0.25, 0.25]$ 米、 $y \in [-0.20, 0.05]$米处，并设有一个较小的中心排除区域；秤在水平面内距离物体至少0.15米。物体和秤的资产变体及平面内朝向均会变化。抓取物体，将其移动到秤的秤盘上，然后释放。评分考查放置情况，而不是数值重量读数。

**SS：印章盖印**（盖印_印章）。印章和一个 $7 \times 7$ 厘米的彩色目标出现在桌子的同一侧，初始间距至少为0.10米。印章外观和目标颜色会变化。将印章移动到指定的彩色目标处并释放。尽管该任务称为盖印，但实际实现的谓词只检查水平对齐和张开的夹爪；并不会单独验证盖印压力或墨印。

**OM：打开微波炉**（打开_微波炉）。一台带铰接门的固定底座微波炉放置在桌子靠后的位置，其型号和位置有所变化。其底部从 $x \in [-0.12, -0.02]$ 米和 $y \in [0.15, 0.20]$ 米中采样。抓住把手并拉动，直至门达到所需的开启角度。无需放入物体、按按钮或进行微波炉操作。

**MP：将订书机移到垫子上**（移动_订书机_垫板）。订书机和彩色垫子（ $11 \times 6$厘米）放置在桌子的同一侧，初始间距至少为0.10米。订书机型号、偏航角和垫子颜色有所变化。抓取订书机，将其移到垫子上，然后释放。释放后的Zero-WAM规则要求订书机与垫子实际接触，并满足水平位置裕度；但不要求满足上游任务的特定四元数条件。

**篮中面包：面包放入篮子**（将_面包_放入篮子）。篮子放置在靠近桌面中心的位置，坐标为 $(0, -0.20)$ 米。面包块从侧边区域随机生成，与篮子及彼此之间留有间隔，且外观和朝向各不相同。场景构造器会尝试放置两块面包；谓词适用于实际生成的每一块面包。将所有面包块移入篮子并释放。若有一块面包留在篮子外，则不满足任务要求。

**EC：杯子放在杯垫上**（将_空_杯放置到杯垫上）。空杯子初始位于左侧或右侧区域，固定的杯垫则放置在更靠近中心的位置，二者初始距离至少为 0.10 米。抓取杯子，将其移至杯垫上，然后释放。得分依据杯子和杯垫上的功能点进行比较；不会单独检测杯子的朝向。

**SB：堆叠三个方块**（将_方块_堆叠成三个）。三个 5厘米立方体分别为红色、绿色和蓝色，初始时分散放置在桌面不同位置，且平面内旋转角度各异。将它们堆成三方块塔，红色在底部、绿色在中间、蓝色在顶部，然后释放两个夹爪。原生谓词会检测方块相对对齐情况和垂直间距；不要求塔位于特定的桌面中心坐标。

## D.2 成功标准

七项任务均采用已发布的 Zero-WAM 评测（表 6）中的二元成功标准。Zero-WAM 为 SS 和 MP 提供了自己的检查，并对其余五项任务采用 RoboTwin 原生检查。长度单位为米。数值 $\Delta x$、 $\Delta y$ 和 $\Delta z$ 表示坐标差。*Both open* 指模拟器对左右夹爪张开的检查。

**表6** 已发布的 Zero-WAM 评估中的二元成功标准。一行中的所有条件都必须满足。

| Task | 成功条件 |
|---|---|
| OS | 物体中心到秤的功能点的水平距离为 $< 0.035$。物体高度为 $> z_{\text{scale}} - 0.01$，且操作夹爪处于打开状态。 |
| SS | 密封件到目标 $\lvert\Delta x\rvert < 0.03$ 和 $\lvert\Delta y\rvert < 0.03$，两者均处于打开状态。 |
| OM | 门关节坐标 $q \geq 0.6 q_{\text{upper}}$，其中 $q_{\text{upper}}$ 为关节上限。 |
| MP | 订书机与垫子 $\lvert\Delta x\rvert < 0.080$和 $\lvert\Delta y\rvert < 0.055$，订书机与垫子的实际接触且两者均处于打开状态。 |
| BB | 每个面包块均满足相对于篮子的 $\lvert\Delta x\rvert < 0.05$、 $\lvert\Delta y\rvert < 0.05$，以及相对于 $z > 0.73 +$ 桌子_z_偏差。两只夹爪均处于打开状态。 |
| EC | 杯子与杯垫功能点水平距离 $< 0.035$和 $\lvert\Delta z\rvert < 0.015$，且两者均处于打开状态。 |
| SB | 对于绿色相对于红色以及蓝色相对于绿色， $\lvert\Delta x\rvert < 0.025$、 $\lvert\Delta y\rvert < 0.025$和 $\lvert\Delta z - 0.05\rvert < 0.012$。两者夹爪处于打开状态。 |

## D.3 EveryTask 的上下文与记录执行

图 12 和 13 展示了七项任务中固定的人类上下文和记录的机器人观测。机器人面板采用时间重置模型在完成 4,000个监督微调步骤后，每项任务的一条成功运行轨迹。参考视频是已发布评测中的固定生成片段。

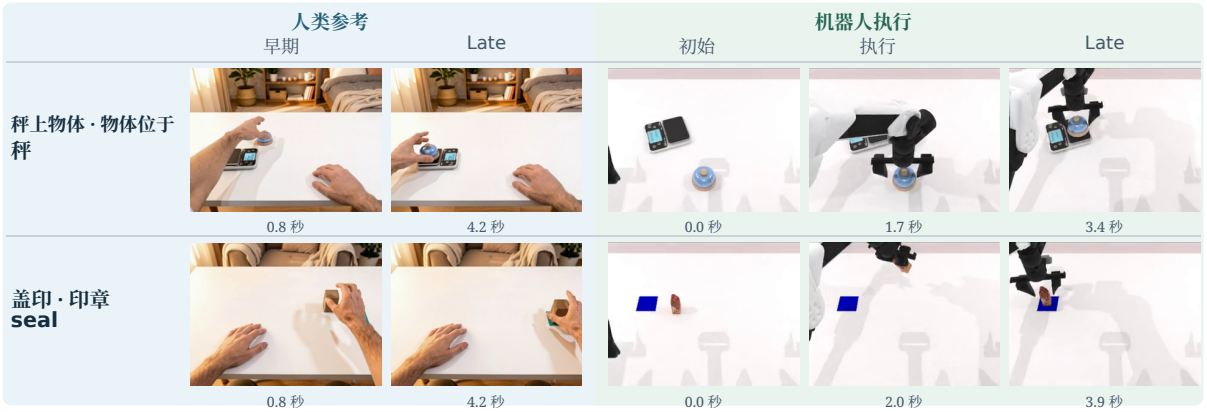



图 12 （第 1 部分，共 2 部分）。




**图 12** RoboTwin 在 OS、SS、OM 和 MP 任务中的环境与交互。两个人类上下文关键帧之后，是时间重置模型在经过 4,000 个 SFT 步数后的初始、中间和晚期机器人观测。机器人面板显示头部摄像头视图。时间戳均相对于各自的源视频。




**图13** RoboTwin 任务环境及其交互，针对 BB、EC 和 SB，使用与图12相同的布局和时间重置模型。

# E 实施与评估详情

## E.1 训练配置与检查点沿袭

真实机器人评估使用未经下游真实机器人SFT的CT检查点。零CT条件使用原始MotuBrain EMA。非零扩展条件使用CT步数4,000的EMA检查点。扩展子集取自由两个同任务示例组成的按来源分层的数据块，并共享全数据验证集划分。仿真分支在CT之后应用多任务RoboTwin SFT，以使模型适应模拟器。

**表7** 真实机器人扩展和RoboTwin实验的训练设置。CT步数预算指扩展运行的步数。RoboTwin SFT模型从对应的CT EMA权重初始化。

| 设置 | CT/真实机器人分支 | RoboTwin SFT分支 |
|---|---|---|
| 训练阶段 | MotuBrain → 持续训练 | MotuBrain →持续训练 → 监督微调 |
| 评估权重 | CT4,000 指数移动平均；MotuBrain EMA 对应 0% | 对应的 SFT4,000 指数移动平均 |
| 阶段长度 | 非零比例对应 4,000 个持续训练步数 | 4,000个监督微调步骤 |
| 全局批次大小 | 1,024 | 128 |
| 监督微调优化器和调度方案 | — | AdamW；学习率 $2 \times 10^{-5}$；100 个预热步数，然后保持恒定 |
| 监督微调微批次 / 随机种子 | — | 2 / 初始化随机种子 6198 |
| 混合训练窗口 | 完整视频中的 13/25/37 个视频帧配方 | 13/25/37 个视频帧 |
| InfoNCE 系数 / 温度 | InfoNCE 配方中的 0.05 / 0.1 | 0.05 / 0.1 |
| 动作统计数据 | CT 专用统计数据；真实机器人动作映射动作映射 | 仅用于 RoboTwin 训练集，逐步 q99 |
| 评估说明 | 空 | 空 |
| 检查点选择 | 已保存的终端指数移动平均 | 已保存的终端指数移动平均 |

**模型配置。** 该模型配置包含 24 个 Transformer 层、4,096 的隐藏维度、32 个注意力头，以及 55 维的统一动作表示。它对头部、左腕

和右腕视角使用加性嵌入。三种 RoPE 条件在训练和推理时采用相同的布局：时间偏移、时间重置和空间偏移。

### E.2 推理设置

**仿真**。 仿真评估客户端将头部和腕部视角合并为一张 $384 \times 464$ 图像，并使用当前观测几何特征。每次动作推理时，上下文包含连续 13 帧原始帧；推理成功后前进 13 帧，上下文耗尽后则重复最后一帧。上下文游标在每个回合开始时重置；客户端不使用未来查询轨迹或学习得到的对齐信号。客户端将视频和动作去噪步数分别设为 50 和 50，视频/动作偏移量设为 5/1，上下文/动作引导设为 5/1，并禁用文本引导。发布的控制器初始执行 16 条指令，之后执行 32 条，并将源动作速率 10 Hz 插值至 50 Hz 的仿真器指令速率。

仿真动作映射采用相机坐标系末端执行器表示和 6D旋转，并使用 RoboTwin训练数据的逐步 q99 统计量。其夹爪通道保留原始 $[0, 100]$ 约定。无参考控制在监督微调期间使用 1,024 的全局批次大小。推理时，该控制接收 0 个上下文令牌，并将上下文引导比例设为 1。

**真实机器人**。 头部和两个腕部 RGB 视角的录制分辨率与帧率见第 C 节。每次推理以 10 Hz 预测 24 个动作（2.4 秒）；自适应时域选择 6–24 步的执行前缀，而无自适应时域控制则执行全部 24 步。上下文窗口包含 13 帧，帧率为 5 Hz，并按已执行前缀的时长向前推进。第 E.3 节说明了同步、注意力提取和引导。10 Hz 指动作航点的频率。

**外部基线**。 正文中的 WAN-Action、LingBot-VA 和 Zero-WAM 条目复现了正文所引 Zero-WAM 论文表 2 中的测量结果。来源采用三个评估随机种子，并针对每个未见任务、每个随机种子运行 100 条轨迹，报告跨随机种子的均值和标准差。我们的评估遵循相同的七项任务划分和已发布的成功检查。结果报告了每个固定检查点在三个评估随机种子上的均值和标准差。

**RoboICL评估**。 我们使用 `xhigh` 推理投入评估 GPT-6 Astra 智能体。我们采用第 D 节和 D.2 节所述的 RoboTwin 任务设置和成功谓词。RoboICL 会在一个回合内保留交互历史，并可重试失败的操作。为保证公平，带星号的条目仅在所有必需操作均首次物理尝试成功时才计为成功。我们在三个随机种子上评估七项任务，每个任务、每个随机种子包含 10 个场景。

### E.3 真实机器人上下文同步与自适应执行

**上下文构建与推进**。 完整的人工演示可作为参考，但每次模型调用都会接收一个局部窗口。在第 $k$ 轮、上下文开始时间为 $\tau_k$ 时，我们以$\tau_k + \{0、0。2、……、2。4\}$ 秒的时间点对源视频进行采样。这会以 5 Hz 提供 13 帧，从第一帧到最后一帧跨度为 2.4 秒。根据时间戳选择最近的源帧，因此播放时间不依赖于源视频的原生帧率。每个回合从 $\tau_0 = 0$ 开始。对于以 10 Hz 执行的 $h_k$ 个动作前缀，我们更新

$$\tau_{k+1} = \tau_k + \frac{h_k}{10}. \tag{5}$$

参考速度乘数为 1。推理、网络和等待延迟不会推进上下文时钟。例如，在观测到上下文 $[0, 2.4]$s 后执行 10 个动作，下一窗口为 $[1.0, 3.4]$ s；执行全部 24 个动作，则为 $[2.4, 4.8]$s。因此，较短的执行会保留完整的上下文窗口，同时使连续窗口之间产生重叠。此对齐方式通过同步执行器累计已下发的动作步数；该执行器会先完成每个前缀，再请求下一次观测。它不会根据编码器估计任务进度，也不会学习上下文与机器人相位对齐。视频结束后的采样会重复最后一帧；示范不会循环播放。开始新回合或切换参考内容时，上下文时钟会重置。

**预测、执行与重新规划**。 每次调用都会接收最新的头部／左腕／右腕 RGB图像、机器人状态和局部人类上下文。模型始终以 10 Hz 预测未来 24 个动作，表示 2.4 s 的动作时间。只有机器人视频潜变量的第一个时间位置以当前观测为条件；未来位置从噪声初始化，因此不会提供真实的未来机器人帧。自适应时域会选择 6–24 个动作作为执行前缀。该前缀完成后，未使用的后缀会被丢弃，并根据新的观测生成全新的 24 动作预测。若不使用自适应时域，每次调用都会执行完整的 24 动作预测。10 Hz 的比率表示模型动作路点的时间间隔；较低层级的执行器可以按更高的比率进行插值。0.6–2.4s 的前缀时长不包括推理延迟，也不是实际时间重规划周期。

**参考注意力统计**。 自适应时域（AH）使用条件前向传播在 50 次更新中的去噪迭代索引 25（从零开始计数）处的注意力。我们选择具有直接视频到动作的连接的 18 个层，每层包含 32 个注意力头；6 个解耦层不包含直接参考键，因此将其排除。对于动作查询 $i$、层 $\ell$ 和注意力头 $b$，令 $P_{ir}^{\ell b}$ 表示可见键 $r$ 上的注意力权重。在 QK归一化、位置编码和条件键拼接之后计算注意力，并在选择参考权重之前对所有可见键执行 softmax。因此，参考会与其他输入竞争注意力质量。令 $\mathcal{R}$ 包含所有参考标记，并令 $\mathcal{R}_g$ 包含潜在时间组 $g$ 中的参考标记。13 帧视频编码为 $G = 4$ 个潜在时间组。在每个注意力头内，定义

$$\mu_i^{\ell b} = \sum_{r \in \mathcal{R}} P_{ir}^{\ell b}, \tag{6}$$

$$q_{ig}^{\ell b} = \frac{\sum_{r \in \mathcal{R}_g} P_{ir}^{\ell b}}{\mu_i^{\ell b}}, \tag{7}$$

$$E_i^{\ell b} = -\frac{\sum_{g=1}^{G} q_{ig}^{\ell b} \log q_{ig}^{\ell b}}{\log G}. \tag{8}$$

在形成参考内部时间分布 $q$之前，先对空间位置求和。其熵衡量的是跨四个潜在时间组的离散程度，而不是跨 13 个原始帧的离散程度。用于选择的统计量是注意力头/层平均值，

$$(m_i, e_i) = \frac{1}{18 \cdot 32} \sum_{\ell, b} (\mu_i^{\ell b}, E_i^{\ell b}). \tag{9}$$

具体而言，先在每个头部中分别计算熵，再取平均值。

**执行长度选择**。 当满足 $m_i < 0.10$ 和 $e_i > 0.86$时，一个动作位置即为截断候选位置。我们要求在同一次预测中出现三个连续的候选位置。若 $j$ 是此类区间的零起始位置，则设定

$$h_k = \mathrm{clip}(j, 6, 24). \tag{10}$$

若没有符合条件的区间，则设定 $h_k = 24$。该规则旨在仅执行候选区间之前的前缀，且至少执行六个动作；因此，若区间起始位置早于位置 6，仍会执行六个动作。自适应时域（AH）是一种确定性启发式方法，并未单独训练时域预测器。注意力质量和熵并非经过校准的任务成功概率或参考跟随置信度分数。

**采样与引导**。 自适应时域（AH）和无自适应时域（AH）推理均使用 50 次欧拉去噪更新。动作时间 $t$从 1 线性递减至 0。对于视频流，我们施加 5 的时间偏移，得到 $t_v = 5t/(1 + 4t)$；每个流均使用各自的相邻时间增量进行更新。每次更新都会评估一个参考条件分支和一个不含参考视频的分支。目标视频流使用参考引导

$$v^v_{\text{guided}} = v^v_{\text{no-ref}} + 5\big(v^v_{\text{ref}} - v^v_{\text{no-ref}}\big). \tag{11}$$

动作流直接取自参考条件分支（动作 CFG 系数为 1），不对两个动作输出进行外推。因此，50 次更新使用 100 次去噪模型前向传播。评估使用空标题。

**AH消融控制**。 无 AH 控制将执行时域固定为 $h_k = 24$ ，而不是选择 6–24 步的前缀。上下文按已执行的时长推进。较短的前缀会增加单位动作时间内的观测获取和重新规划次数。

### E.4 配对数据集构建与划分检查

**回合级划分**。 一个目标机器人回合可以与多个人类视频配对。我们将所有共享同一目标回合的配对分配到同一划分中，并确认训练集和验证集之间没有重复的目标回合标识符。

**同步人类重演**。 参与者先练习跟随源机器人视频，然后一边观看视频回放，一边录制并同步操作动作。参与者采用与头部摄像头相匹配的视角，用手模仿夹爪的抓取和释放动作，同时避免进行平行夹爪无法复现的灵巧手指操作。所有人类演示仅记录桌面手部和前臂操作，不包含面部或个人身份信息，且均在征得参与者同意后采集。

**关键帧编辑与视频生成**。 源机器人视频标注（包括由 Qwen3.8 生成的描述）提供原子动作条件。使用 `gpt-image-2` 和 HaGRID 人手外观参考编辑选定的源帧。编辑保留源场景、物体、视角和接触关系，同时将机器人操作器替换为人手和前臂。经过编辑的帧在作为外观参考用于 Seedance 2.5 或 Wan 3.0 进行视频编辑之前，会先通过质量检查。完整的源机器人片段与动作描述及场景保持指令共同提供运动参考。不会向生成器提供明确的机器人臂分割掩码、深度图或边缘图。经认可的人类视频片段保留原始机器人观测和动作作为监督信号。

**跨具身动作监督**。 Astribot 和 AgiBot 的录制数据都提供配对视频–动作训练样本。AgiBot 样本在共享动作表示下保留各自录制的动作，并通过维度有效性掩码标明可用的动作分量。在持续训练期间，这些样本提供视频和动作监督；真实机器人部署则在 Astribot S1 上进行评估。

**VLM筛选**。 我们在每个片段等时长区间的中点处采样帧，并按时间顺序将这些帧提交给视觉语言模型（VLM）。进行成对比较时，来源帧和生成帧分别置于两行中，每一列对应两个片段中相同的相对时间点。

视觉质量检查会检验可见机器人操作器是否已被完全替换、是否不存在残留的机器人几何结构、是否呈现符合解剖结构的手部和前臂外观，以及场景是否稳定、运动是否连续。我们会区分来源中原本就存在的任务工具与机器人残留物。即使表面外观类似皮肤，呈肤色的刚性机器人连杆或夹爪形状也无法通过检查。

进行源对齐检查时，验证器会接收成对帧和源动作描述，并检查动作身份、左右角色、主要阶段的时间安排、轨迹、物体状态，以及是否存在额外操作。如果主要操作和最终物体状态一致，则允许手部姿势不同、放置后有偶然接触，以及释放或收回时机略有差异。更换目标物体、反转运输方向，或在源动作结束后继续操作物体，均无法通过此项检查。验证器会为每项检查返回通过/失败判定。只有同时通过两项检查的生成片段才会被保留。