arXiv:2602.23648v1 [cs.RO] 27 Feb 2026

FAVLA：用于高接触机器人操作的力自适应快慢 VLA 模型

Yao Li [*1] **Peiyuan Tang** [*2] **Wuyang Zhang** [1] **Chengyang Zhu** [3] **Yifan Duan** [1] **Weikai Shi** [1] **Xiaodong Zhang** [1] **Zijiang Yang** [1] **Jianmin Ji** [1] **Yanyong Zhang** [1]

摘要

力/力矩反馈可以显著提升视觉-语言-动作（VLA）模型在富接触操作中的表现，但现有大多数方法都以单一工作频率融合所有模态。这种设计忽略了真实机器人传感器采样率的不匹配，迫使对反应式校正所需的高频接触线索进行下采样。再结合常见的视觉语言模型-动作专家（AE）流水线——它们在代价高昂的视觉语言模型更新之间主要以开环方式执行动作块——统一频率融合往往会对冲击、粘滑和力尖峰产生滞后响应。我们提出 FAVLA，一种将慢速感知规划与快速接触感知控制解耦的力自适应快慢视觉-语言-动作。FAVLA 以固定低频运行慢速视觉语言模型，对模态进行编码以产生潜在表征，并预测近期力变化。随后，快速动作专家以可变高频执行，并以最新力序列数据为条件生成反应式动作。我们进一步引入力适配器，将高频力特征注入多个动作专家层，并根据视觉语言模型预测的力变化自适应调度动作专家的执行频率。在高接触任务上的大量实验表明，FAVLA 显著优于基线方法，取得了更优的反应能力和成功率，尤其是在操作过程中接触力更小的情况下。

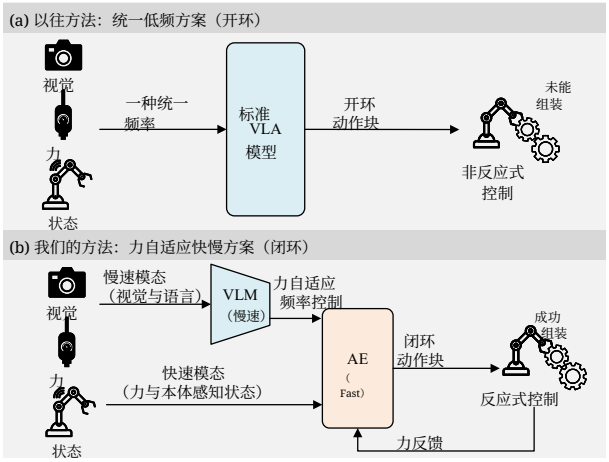


图 *1.* **视觉-语言-动作模型以往的统一频率方案与我们的力自适应快慢方案对比**。(a) 以往方法以统一的输入频率处理视觉、力和本体感知状态，并生成开环动作块，无法及时利用高频力反馈。(b) 我们的方案采用力自适应快慢方案。视觉语言模型以较低频率运行，对视觉和语言上下文进行编码；动作专家则利用最新的力数据，以力自适应的高频率运行，生成闭环动作块，从而实现反应式机器人控制。

1. 引言

视觉-语言-动作（VLA）模型近期已逐渐应用于通用机器人操作，其通过将大型视觉语言模型（VLM）的语义泛化能力与连续控制策略相结合（通常由动作专家实现）。借助大规模多任务机器人预训练，π_0类 VLA 模型（Black 等人，2024；2025 年；Pertsch 等人，2025 年）已展现出利用视觉、语言和本体感觉状态在不同机器人及任务间迁移的良好前景。然而，在富接触操作（如插入和紧公差装配）中的可靠表现，仍比自由空间抓放更难实现。在这些场景中，成功不仅取决于视觉对齐，还取决于接触条件变化时能否迅速进行运动修正，包括应对仅凭视觉难以充分观测的突发事件（冲击、粘滑、卡滞）。

一种自然的做法是将包括力/力矩数据在内的交互信号纳入视觉-语言-动作模型的输入。近期研究表明，这种做法能够提升高接触任务的成功率

[1]中国科学技术大学，中国合肥 [2]西安交通大学，中国西安 [3]中南大学，中国长沙。通讯作者：Wuyang Zhang <wuyangz@ustc.edu.cn>、Yanyong Zhang <yanyongz@ustc.edu.cn>。

预印本。*2026*年*3*月*2*日。

通过向视觉-语言-动作模型添加力/力矩反馈（Yu 等人，2025 年；Zhang 等人，2025b；Shahidzadeh 等人，2025 年）或将其加入动作策略（Xue 等人，2025 年；Chen 等人，2025b）。然而，大多数现有方法都在单一运行频率下，通过单一令牌流融合多种模态（例如，将力令牌与视觉令牌拼接）。实际上，这仍留下两个尚未解决的挑战。首先，各模态的词元化通常不均衡：视觉产生的词元远多于力数据，因而可能使注意力偏离稀疏的力线索，即使这些线索对接触调节至关重要。其次，传感器的采样率并不匹配（例如，15 Hz 摄像头与100–1000 Hz 力/扭矩传感器）。在标准的统一频率处理流程中，高频力数据会被降采样至与频率最低的模态一致，从而丢弃那些能够指示故障并要求立即修正的响应信号信息。

这种频率不匹配是 VLM–AE 控制模式中的一个重要影响因素。主流的 VLM-AE 范式采用低频 VLM 推理和开环动作块执行策略来实现机器人控制，因此难以在执行过程中针对状态/力变化及时调整动作。这一问题在工业富接触场景中尤为关键，例如高精度装配；延迟调整动作可能导致操作力过大和部件损坏。这些局限目前阻碍了基于力的视觉-语言-动作模型在实际工业操作中的部署。

本文认为，接触丰富型视觉-语言-动作控制应明确遵循快慢感知与控制原则：语义场景理解天然是“慢”的，而接触反馈和纠正控制则应当是“快”的。我们通过 FAVLA 将这一理念付诸实践。FAVLA 是一种力自适应快慢视觉-语言-动作模型，它将语义推理与力响应执行解耦，同时保留端到端学习方案。具体而言，我们首先提出一种 力注入式快慢视觉-语言-动作架构，将“快”和“慢”两种模态整合到端到端视觉-语言-动作模型中：(i) 一个以固定低频运行的慢速视觉-语言模型主干网络，用于将视觉、语言和历史力测量值整合为潜在上下文表示；(ii) 一个以更高频率运行的快速动作专家，根据最新的本体感知状态和力测量值，以闭环方式优化动作块。同时，为了让动作专家能够直接根据接触力数据调整动作（而非仅将其作为额外词元附加），我们引入力适配器，将最新的力信息注入多个动作专家层。其次，为了实现响应式动作调整，我们进一步提出一种 力自适应快慢推理策略。与视觉语言模型不同，动作专家以可变高频运行，并根据视觉语言模型预测的未来力变化进行自适应调度。这一设计使 FAVLA 在自由空间运动期间保持低频语义推理和动作输出，并且——

在自由空间运动期间保持低频语义推理和动作输出，而在接触转换即将发生或正在发生时，则提升至高频进行动作调整。因此，该系统能够响应式地调整动作，避免工业应用中产生过大的操作力。

我们在一系列富接触操作任务上评估了 FAVLA，这些任务要求精确的力调节和快速响应以调整动作。结果表明，力自适应快慢机制在保持较低峰值接触力的同时，显著提高了成功率和反应能力。与最强的基线方法相比，FAVLA 将任务成功率提高了 13.8%，平均成功率达到 80.8%，验证了其在高精度工业操作中的有效性。

总而言之，本文的贡献如下：

- 我们提出了一种力自适应快慢视觉-语言-动作框架，即 FAVLA，它将慢速语义模态（视觉、语言和力历史）与快速交互信号（最新力数据）融合，用于闭环接触密集型控制。
- 我们提出了一种力注入动作专家设计，其中包含一个力适配器，用于整合高频力反馈，从而能够直接进行力条件校正，修正动作块。
- 我们提出了一种力自适应快慢推理策略：慢速视觉语言模型预测近期力变化，以此调度快速动作专家的执行频率，从而提高接触过程中的响应速度，同时减小操作力。
- 在高接触任务上的大量真实世界实验表明，与力感知和纯视觉基线方法相比，我们的方法提高了成功率，并降低了峰值接触力。

2. 相关工作

视觉-语言-动作模型。如今，视觉-语言-动作模型已成为一种强大的范式，通过大规模多模态预训练学习通用机器人策略（Zhang 等，2025a）。自回归方法（Brohan 等，2022；Zitkovich 等，2023；Kim 等，2024）将控制表述为对离散化动作进行下一词元预测，从而能够利用大规模异构数据集进行可扩展学习。在此基础上，生成式视觉-语言-动作模型（Black 等，2024；2025；Pertsch 等，2025；Zhai 等，2025）通过流匹配或扩散式目标对连续动作轨迹进行建模，以解决动作离散化问题。近年来，分层视觉-语言-动作架构（Chen 等，2025a；Zhang 等，2024；Wen 等，2025）提出了将高层推理与低层执行解耦的新范式，从而提升稳定性和效率。受此启发

，我们利用大规模预训练视觉-语言-动作模型的优势，并为高接触机器人操作采用快慢策略。

接触丰富操作模型。近期关于富接触操作的研究，通过明确纳入视觉之外的物理交互信号，扩展了基于学习的策略（Feng 等人，2024；Liu 等人，2025；He 等人，2025）。力感知策略（Xue 等人，2025；Chen 等人，2025b）引入慢-快扩散架构，利用力或与力相关的反馈，在接触期间实现快速纠正行为，同时保持稳定的长时程执行。后续研究（Yu 等人，2025；Zhang 等人，2025b；Cheng 等人，2025）将力或力矩信号整合到视觉-语言-动作框架中，使预训练的 VLA 能够根据接触状态调整执行策略。另有一类研究倾向于采用基于触觉的方法（Huang 等人，2025a；Li 等人，2025；Huang 等人，2025b；Bi 等人，2025），通过纳入视触觉感知，提供细粒度的接触感知和触觉感知执行。然而，现有方法通常假设运行频率统一，这会舍弃高频力线索，并限制反应式接触控制。我们的方法提出了一种力自适应快慢框架，在不同时间尺度上解耦语义推理与力反馈，从而提升控制响应性。

总之，当前典型的基于力的视觉-语言-动作模型将所有模态统一为单一输入频率（例如，对高频模态进行下采样以匹配低频模态），这会丢弃高频接触线索，并可能导致开环行为和力尖峰。与采用统一运行频率的主流VLM-AE流水线不同，FAVLA引入力注入快慢融合策略，可根据未来接触力自适应地调整动作。同时，视觉语言模型以固定的低频率运行，利用视觉、语言和力历史数据预测未来的接触力趋势；而动作专家则以更高且可变的频率运行，利用最新的高频力信号进行反应式动作校正。凭借这些设计，FAVLA能够有效减少力尖峰，并提高高精度工业接触任务中的安全性和成功率。

3. 方法

我们提出了 FAVLA，这是一种力自适应视觉-语言-动作（VLA）模型，采用快慢融合策略：视觉和语言提供缓慢变化的语义上下文，而本体感觉和力则需要高频更新，以调节接触并响应交互瞬变。第 3.1 节将需要响应式控制的高接触任务形式化，并回顾标准的 $\pi0$ (Black 等人，2024) 框架以及我们的 FAVLA 框架。第 3.2 节介绍注入力的快慢 VLA 架构。第 3.3节介绍一种力自适应推理策略，根据预测的未来力变化，以可变频率调度动作专家。

3.1. 预备知识

问题定义。我们旨在构建一种快慢视觉-语言-动作模型，利用不同频率的多模态观测完成高精度工业操作任务。这些任务要求处理接触力较小的工业产品，并根据机器人的观测实现反应式闭环控制，尤其是在接触力增大时。在时间步 t，我们将机器人的观测表示如下：

$$O_t = \{\mathcal{I}_t^{(k)}\} \cup \{\mathbf{s}_t\} \cup \{\mathbf{f}_{t-\tau+1:t}, \mathbf{f'}_{t-\tau+1:t}\}, \quad (1)$$

其中 $\mathcal{I}_t^{(k)}$ 表示第 k帧RGB图像（在我们的设置中，分别来自外部相机和腕部相机），$\mathbf{s}_t \in \mathbb{R}^7$ 表示六自由度TCP（末端执行器的工具中心点）位姿和夹爪宽度的本体感知状态，$\mathbf{f}_{t-\tau+1:t} \in \mathbb{R}^{\tau\times 6}$ 表示一个时间窗口内的历史力数据序列，其中包含长度为 τ的时间窗口内的六轴力测量值；$\mathbf{f'}_{t-\tau+1:t} \in \mathbb{R}^{\tau\times 6}$表示时间 t 时最新的力数据序列，其时间窗口长度同为 τ。在我们的设置中，我们在末端执行器上安装了一个六轴力传感器，同时，$\mathbf{f}_t = \{f_x^t, f_y^t, f_z^t, m_x^t, m_y^t, m_z^t\}$包含末端执行器的三维力和力矩分量。给定语言操作指令 L 和观测 O_t，我们的目标是学习端到端视觉-语言-动作模型 $\pi(A_t|O_t, L)$，以最大化真实操作轨迹的似然，并输出低层级动作块 $A_t = \{a_t, a_{t+1}, ..., a_{t+H-1}\}$。与此同时，在观测中，$\mathbf{s}_t$ 和$\mathbf{f'}_{t-\tau+1:t}$ 以较高频率更新，而 $\mathcal{I}_t^{(k)}$ 和$\mathbf{f}_{t-\tau+1:t}$ 以较低频率更新。我们希望视觉-语言-动作模型能够根据高频输入，尤其是最新的接触力，反应式地调整动作块。

框架 $\pi0$以及我们的 FAVLA。FAVLA 是一种基于 $\pi0$ (Black 等人，2024) 框架、面向富接触操作任务的力自适应快慢视觉-语言-动作模型。它由一个较大的视觉-语言模型主干网络和 PaliGemma 的一个较小动作专家组成（Beyer等人，2024）。如图 2所示，FAVLA 将不同频率的多模态输入分别输入视觉语言模型和动作专家，然后使用条件流匹配模型生成动作轨迹。不同模态被编码为词元，并以不同频率输入 FAVLA。同时，较小的动作专家具有更高的推理频率，其频率会根据预测的未来接触力变化进行自适应调整。

3.2. 力注入式快慢视觉-语言-动作模型架构

FAVLA 旨在融合采样率各异的多模态感官数据。我们将视觉-语言-动作模型分解为：(i) 慢速 VLM 主干网络，利用视觉、语言和低频力历史序列进行语义场景与指令条件化；

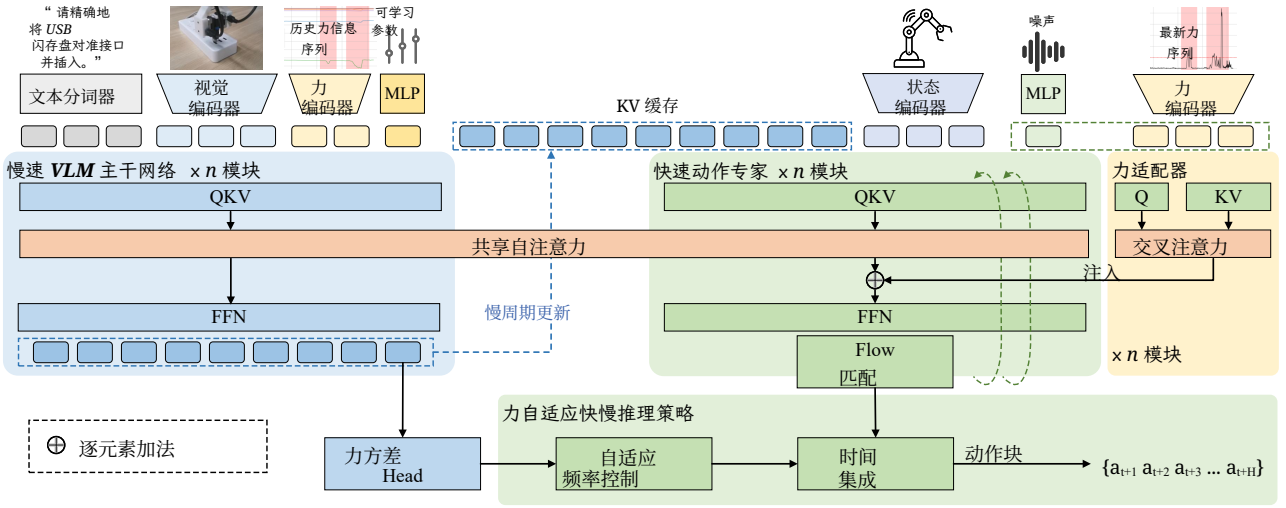


图 *2.* **FAVLA 概览**。FAVLA 集成了一个大型慢速 VLM 主干网络，用于语义推理；同时配备一个较小的快速动作专家，用于响应式动作控制。模型以不同频率处理多模态输入，并通过条件流匹配生成动作轨迹。值得注意的是，动作专家的推理频率会由力自适应快慢推理策略自适应调整，从而确保操作过程中的接触力较小。

以及 (ii) 快速动作专家，利用最新的高频力测量值执行力响应控制。慢速主干网络为快速模块生成可复用上下文（记为 KV 缓存），而快速动作专家则注入力信息，以在多个层级上优化动作。

慢速 VLM 主干网络。 慢速组件是一个整合视觉、语言和历史力数据的 VLM Transformer。来自外部摄像头和腕部相机的 RGB 图像由 SigLIP 图像编码器编码为图像块令牌，语言指令则由 Gemma 分词器嵌入。此外，我们采用轻量级时序卷积网络（TCN）分词器对历史力数据 $\mathbf{f}_{t-\tau+1:t}$ 进行编码，生成历史力令牌 $\mathbf{z}_f$。具体而言，我们首先应用 1×1 时序卷积，将每个时间步的力向量投影为 d_f 维特征，随后接入一组带残差连接的因果膨胀时序块。每个模块使用两个带膨胀操作和层归一化的因果一维卷积。最后，我们将所得序列下采样为 N_f 个令牌，并生成历史力令牌 $\mathbf{z}_f \in \mathbb{R}^{N_f\times d_f}$.

我们将视觉令牌、语言令牌和力令牌 $\mathbf{z}_f$ 拼接起来，作为慢速 VLM 主干网络的前缀令牌，从而实现跨模态的完整自注意力。主干网络输出最终的视觉-语言-力（VLF）令牌 $\mathbf{H}_{\text{VLF}} \in \mathbb{R}^{S_{\text{pre}}\times d}$，其中 S_{pre} 是前缀令牌的数量，d 是模型宽度。为供快速模块重复使用，我们在每一层缓存前缀令牌的中间层键/值状态 $\mathbf{K}^{(\ell)} \in \mathbb{R}^{S_{\text{pre}}\times n_{kv}\times d_h}$，和 $\mathbf{V}^{(\ell)} \in \mathbb{R}^{S_{\text{pre}}\times n_{kv}\times d_h}$，其中 n_{kv} 是键/值头的数量，d_h 是头维度。

快速力注入动作专家。*Fast*组件是一个小型 Transformer，通过高频反馈来优化动作块。在时间 t，它通过流匹配生成动作块，条件是最新力序列 $\mathbf{f'}_{t-\tau+1:t}$ 和机器人状态 $\mathbf{s}_t$，以及慢速更新的 VLM 键/值缓存 $\{\mathbf{K}^{(\ell)}, \mathbf{V}^{(\ell)}\}$。同时，我们不将力序列输入动作专家的输入序列（因为这可能会稀释稀疏的力线索），而是通过力适配器将其注入不同层中。具体而言，我们对 $\mathbf{f'}_{t-\tau+1:t}$ 应用相同的 TCN 分词器，以获得最新力令牌 $\mathbf{z'}_f$。随后，在每个 Transformer 层中，带噪动作令牌 $\mathbf{z}_a$ 在与键/值缓存进行自注意力计算后，会对 $\mathbf{z'}_f$ 执行交叉注意力，并以加法方式注入由此得到的力条件更新：

$$\mathbf{z}'_a = \mathbf{z}_a + \text{Attn}(\mathbf{Q}_a, \mathbf{K}_f, \mathbf{V}_f), \tag{2}$$

其中，$(\mathbf{K}_f, \mathbf{V}_f)$ 是通过将 $\mathbf{z'}_f$ 投影到力键/值而获得的，而 $\mathbf{Q}_a$ 是带噪动作查询。这种力注入使动作专家能够根据高频接触信号直接调整动作块。

力方差头部。基于这一快慢速 VLA 模型架构，我们添加了一个辅助力方差头，用于预测操作过程中接触力未来的波动性。它指导如何动态调整第 3.3 节所述的推理频率。我们从 $\mathbf{H}_{\text{VLF}}$ 中选取 VLF 令牌，该令牌对应于用于方差预测的可学习输入令牌。随后，我们将其输入多层感知机头，生成一个标量预测值。预测目标是对六维力/力矩信号进行指数移动平均滤波后得到的加权方差，计算窗口为由动作块大小定义的短期未来窗口。

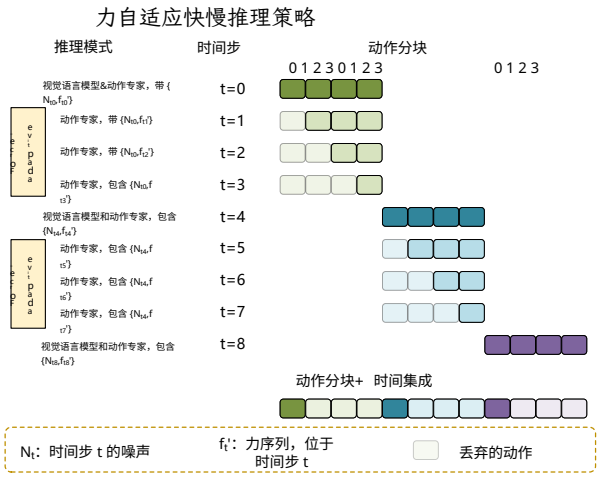


图*3.* **我们的力自适应快慢推理策略**。 快速动作专家在每个动作块内运行多次，并以实时力为条件。在每个周期中，我们固定采样噪声和机器人状态，然后对重叠动作块进行时间集成。动作专家执行频率根据视觉语言模型预测的力方差进行自适应设置。

接下来，我们将介绍如何计算训练中使用的力方差监督标签。给定来自数据集的未来力序列 $f_{t:t+W-1} \in \mathbb{R}^{W\times 6}$，我们将原始波动值计算为加权方差：

$$\nu_t \;=\; \sum_{j=1}^{6} w_j \,\mathrm{Var}\big(\mathbf{f}^{(j)}_{t:t+W-1}\big). \tag{3}$$

随后，我们使用指数移动平均（EMA）平滑算法对时间序列进行平滑，以降低测量噪声：

$$\bar{\nu}_t \;=\; \mathrm{EMA}(\nu_{\le t};\,\alpha). \tag{4}$$

之后，我们使用 $\tilde{\nu}_t =$ 的 $tanh$ 函数 $\tanh\left(\frac{\sqrt{\bar{\nu}_t}}{\sigma}\right)$ 对 $\bar{\nu}_t$ 进行归一化操作，然后训练力方差头，以回归归一化后的 $\tilde{\nu}_t$。监督标签在数据集预处理中生成。

训练目标。 我们通过对动作生成损失和辅助方差预测损失进行加权求和，联合优化慢速视觉语言模型、快速动作专家和力方差预测器：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{action}} + \lambda \mathcal{L}_{\text{var}}, \tag{5}$$

其中，λ 是用于平衡辅助目标的超参数。我们在所有实验中均将 $\lambda = 0.1$ 设为固定值。

3.3. 力自适应快慢推理策略

为了在动作块内实现高频闭环控制，我们提出一种力自适应Fast–Slow Inference策略。其核心思想是将低频视觉推理与高频力反馈解耦，从而发挥二者的互补优势。

在每个周期中，视觉语言模型执行一次前向传播，以处理多模态观测并构建键值缓存，从而提供语义上下文。动作专家根据键值缓存和最新的力观测，多次运行以生成高频动作，进而实现闭环反应式控制。然而，这种简单的推理策略存在两个主要局限：（1）固定频率无法适应任务阶段。自由空间运动和接触密集阶段对推理频率的需求截然不同，但静态调度却一视同仁。（2）缺乏动作块一致性。由于模型具有随机性，同一动作块内独立的动作专家推理可能产生不一致的动作，导致运动抖动且不稳定。针对这些局限，我们在下文提出两种相应的设计。

自适应推理频率。 我们不使用固定的推理频率，而是将其设为随接触状态变化的动态参数。当接触即将发生或正在进行时，我们更频繁地运行动作专家，以实现更好的反应式控制；在自由空间运动期间，则降低运行频率以节省计算量。

具体而言，我们使用视觉语言模型的预测力方差 $\tilde{\nu}_t \in [0,1]$ 来衡量下一周期接触动力学的活跃程度。给定最大频率比$N_{\max}$，我们将动作专家执行的次数安排为：

$$n_t = \max\left(1, \lceil \tilde{\nu}_t \cdot N_{\max} \rceil\right). \tag{6}$$

当 $\tilde{\nu}_t$ 较小时，系统以较低频率（$n_t = 1$）运行。随着 $\tilde{\nu}_t$ 接近1，系统会安排更多动作专家推理，从而在发生接触时提供更快的响应。

一致动作集成。 为应对重复动作专家推理的随机性所造成的动作块内部不一致性，我们采用两种策略。首先，我们将采样噪声 $\boldsymbol{\epsilon}$ 固定在同一视觉周期内的所有动作专家调用中，确保所有推理共享同一随机样本。其次，如图3所示，我们应用时间集成（Zhao等人，2023），以聚合跨视觉周期的重叠动作块预测，从而在各周期间产生更平滑的动作执行。

4. 实验

在本节中，我们针对多种真实世界接触密集任务，对我们所提出的方法进行了全面评估。我们评估了三个关键方面：（1）与强基线方法相比，我们方法的有效性；（2）验证各模块贡献的消融实验；（3）可视化结果与分析。

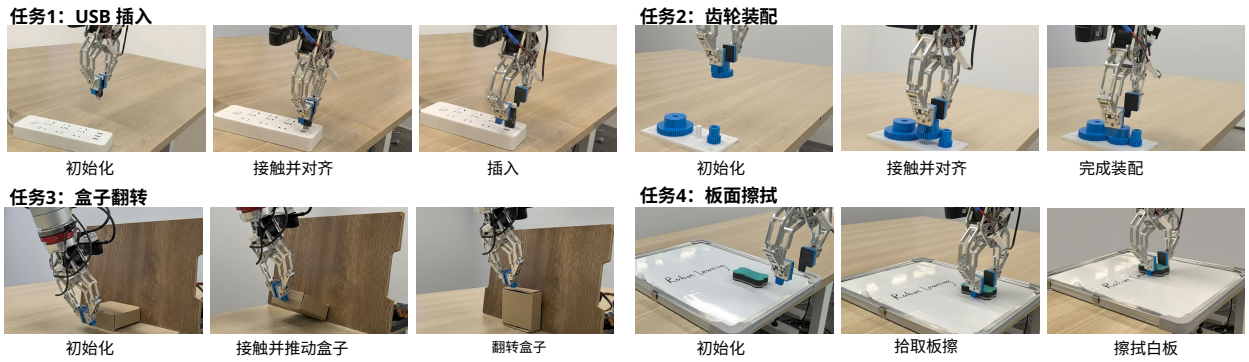
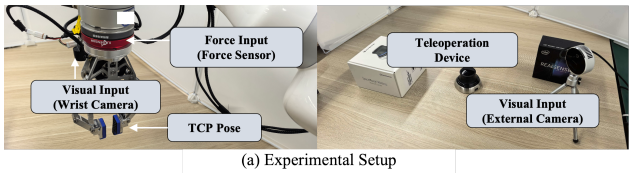
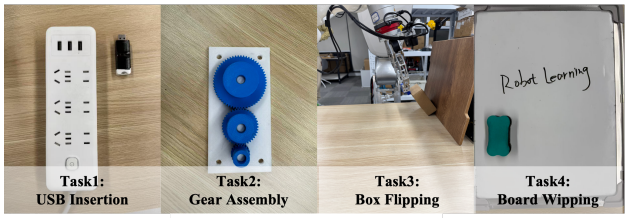


图 *4.* 真实世界任务场景与执行阶段。任务包括（1）USB 插入、（2）齿轮装配、（3）盒子翻转和（4）板面擦拭。这些任务带来了多样化的挑战：任务 1 和 2 需要毫米级精度才能成功对准，而任务 3 和 4 则需要实时力调节，以应对不断变化的接触动力学。



(a) Experimental Setup



(b) Experimental Tasks

图 *5.* 机器人操作实验装置。(a) 硬件配置概览，包括遥操作设备（3D Space Mouse）、6D 力传感器和视觉传感器，以及机器人夹爪。(b) 实验任务集。

4.1. 实验设置

硬件平台。我们使用 Monte 机器人，这是一款配备 7 自由度 X-ARM 机械臂的双臂全身机器人。每条机械臂均在末端执行器法兰处配备腕部安装的 RGB-D 相机和六轴力/扭矩传感器。我们使用腕部相机和外部固定相机采集视觉观测，如图5。

实验任务。为评估所提出方法的有效性，我们针对四项任务开展实验：USB 插入、齿轮装配、箱体翻转和白板擦拭，如图所示。4。这些任务可分为两组：前两项是高精度任务，需要达到毫米级精度并提供实时力反馈，以防因用力过大而导致任务失败；后两项则是涉及动态交互和力反馈的高接触任务。我们为每项高精度任务采集了 80 条示范轨迹，为每项高接触任务采集了 50 条轨迹。人工示范数据通过使用 3D SpaceMouse 进行遥操作采集，所有轨迹均经过同步处理，并降采样至统一的 30Hz 帧率。

对比基线方法。 我们将本方法与以下强基线方法进行比较： π_0 (Black 等人，2024)， π_0 + Force、TA-VLA（Zhang 等人， 2025b）以及 ForceVLA（Yu 等人，2025）。 π_0 是一种基于流匹配的视觉-语言-动作模型，在大规模数据上预训练，在真实世界任务中表现出很强的泛化能力。 π_0 + Force 通过将力数据转换为词元并与输入序列拼接，以融入力信息。TA-VLA 引入了用于未来力预测的辅助目标，以增强模型对力信息的注意力。ForceVLA 采用混合专家（MoE）（Shazeer 等人，2017）架构，从力和图像模态中学习特征。为进行公平比较，所有方法均使用 LoRA（Hu 等人，2022）在我们收集的数据集上进行微调。

评估指标。 我们使用任务成功率来评估模型性能。对于每个操作任务，我们基于任务完成情况建立二元成功标准（例如，USB 完全插入、齿轮正确对齐）。成功率计算为成功试验次数与总试验次数之比，从而直接衡量策略性能。评估分别在 USB 插入的 25 次试验、齿轮装配的 30 次试验、盒子翻转的 20 次试验以及板面擦拭的 10 次试验上进行。我们还使用平均峰值接触力（N）来评估操作过程中的接触力。

实现细节。 我们遵循 π_0 架构，其中视觉-语言-动作由一个视觉语言模型和一个动作专家组成。我们使用 π_0 权重初始化模型，并应用 LoRA（Hu 等人，2022）对视觉语言模型和动作专家进行微调。LoRA 参数配置为$r = 16$, $\alpha = 16$ 用于视觉语言模型； $r = 32$, $\alpha = 32$ 用于动作专家，并应用于注意力和前馈层。我们对所有模型进行了 30,000 次迭代训练，

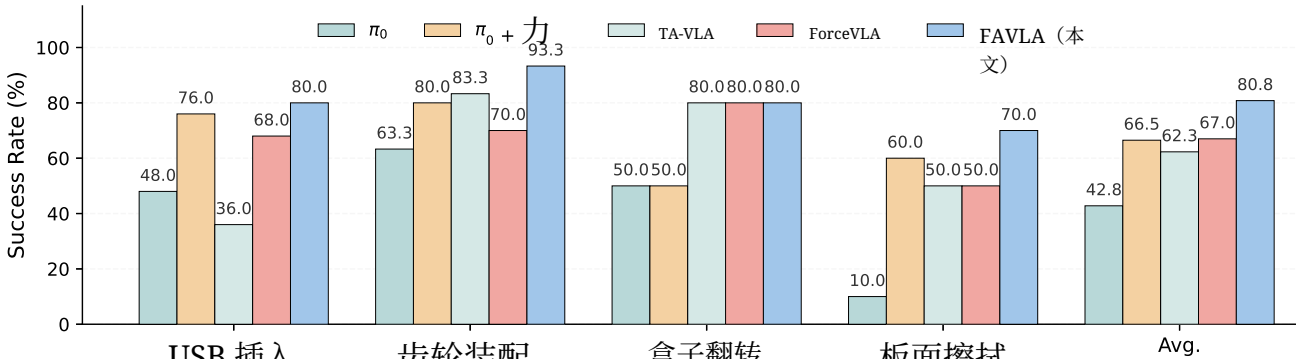


图 *6.* 我们提出的方法与基线方法在四项真实世界任务上的成功率对比。我们的方法在四项高接触任务上显著优于所有基线方法。

学习率为 2.5×10^{-5}，采用余弦衰减调度。批次大小设为 8，训练在配备 80GB 内存的 Nvidia A100 GPU 上进行。

4.2. 评估结果

总体性能。 如图6所示，FAVLA 在所有高接触操作任务上均取得最佳性能。我们的方法平均成功率为 80.8%，比纯视觉 π_0 基线方法高出 38.0 个百分点，并比最佳基线方法 ForceVLA 高出 13.8 个百分点。这些结果凸显了我们方法的有效性。

我们观察到，齿轮装配（93.3%）和板面擦拭（70.0 %）的性能提升尤为显著。这归因于这些任务对力非常敏感。值得注意的是，齿轮装配对过大的力十分敏感，即使稍微用力过度，也可能触发安全阈值并导致失败；而板面擦拭则要求在整个运动过程中保持稳定的接触力。我们提出的力自适应快慢推理机制通过提供力闭环控制来解决这一问题，使系统能够根据高频力反馈响应式地调整动作。

力控制能力。 我们通过测量平均峰值接触力，进一步验证力控制能力。该指标表示每次试验中末端执行器的最大接触力，并对每项任务的所有试验取平均值。我们选择箱体翻转和齿轮装配任务进行评估，因为在操作力过大时，齿轮和纸箱很容易损坏。如表 1所示，FAVLA 将接触力维持在安全且有效的范围内，在齿轮装配任务和箱体翻转任务中的峰值接触力分别为 7.7 N 和 9.9 N；与纯视觉基线相比，分别降低了 4.3 N 和 2.3 N。相比之下，诸如 π_0 和 π_0 + Force 等基线方法经常施加过大的力来完成任务，可能会损坏物体或机器人本身。峰值接触力显著降低，证明了我们的自适应快慢推理策略有效；该策略通过高频闭环动作控制实现精确的力调节。

表 *1.* 高接触任务中的平均峰值接触力（N）。数值越低，表示操作越安全、顺应性越好。

方法	齿轮装配	箱体翻转
π_0	12.0 N	12.2 N
π_0 +力	13.0 N	13.8 N
TA-VLA	11.3 N	10.0 牛
ForceVLA	10.9 牛	12.4 牛
FAVLA（我们的方法）	**7.7 牛**	**9.9 牛**

表*2.* **对所提出组件的消融研究**。 我们通过逐步将各组件添加到纯视觉基线中，验证每个组件的有效性。

配置	盒子翻转	板面擦拭
纯视觉	50%	10%
+ 力注入自编码器	65%	60%
+ 力方差预测	70%	60%
+ 力自适应推理	**80%**	**70%**

4.3. 消融实验

对所提出组件的消融研究。 为验证我们的架构设计，如表 2 所示，我们通过将每个提出的组件逐步集成到纯视觉基线中，开展消融研究。纯视觉基线的能力有限，在盒子翻转和白板擦拭任务上的成功率分别仅为 50% 和 10%。向动作专家注入力可将性能提升至 65 % 和 60%，证明了引入力模态的有效性。引入力方差预测后，盒子翻转的成功率进一步提升至

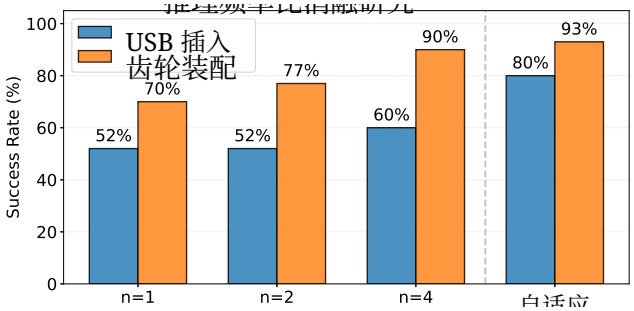


图 7. **频率比消融研究**。 静态频率比（每个视觉语言模型步骤对应 n 个动作专家步骤）与我们提出的自适应频率机制的性能比较。在两项任务中，我们的自适应方法均显著优于所有静态频率设置。

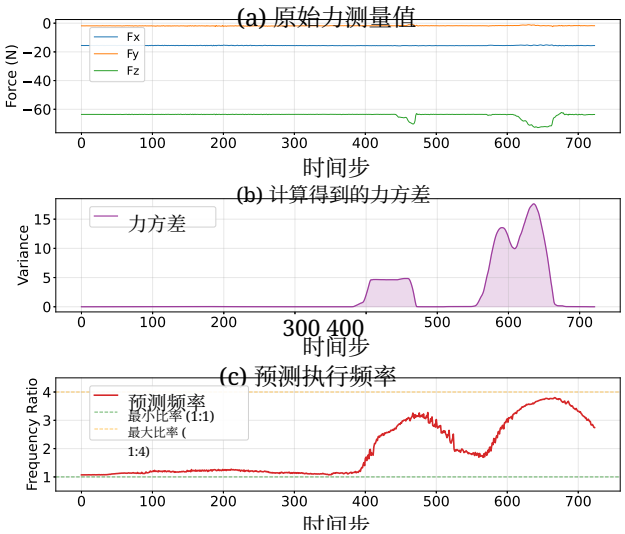


图 *8.* **自适应频率预测可视化**。 (a) 原始力测量值。(b) 计算得到的力方差用于识别环境动态。(c) 模型预测的频率比，表明在更可能发生接触的区域，执行频率有所提高。

70%，这表明，对未来接触力的不确定性进行建模有助于模型更好地理解复杂的操作物理学，从而生成动作。最后，加入带有自适应频率机制的力方差预测模块后，性能提升至 80% 和 70%，验证了其有效性。

力自适应频率调节消融研究。 为评估力自适应频率机制的有效性，我们在 USB 插入和齿轮装配两项操作任务上开展了消融实验。我们比较了三种推理策略：(1) 采用固定频率，固定比率为 $n = 1$、2、4；(2) 我们提出的力自适应频率机制。如图 7所示，结果表明，在固定频率策略中，成功率随着执行频率 n 的提高而持续提升。这说明，较高的推理频率通常能让高接触任务获得更好的响应能力。值得注意的是，我们提出的力自适应策略始终优于所有固定频率方法，在齿轮装配任务中取得了 93% 的成功率，并将 USB 插入任务的成功率显著提升至 80%。这表明，根据实时力需求动态调整频率，比始终保持较高的固定频率更有效，因为这样可以避免在非接触阶段频繁切换不同的动作块，从而确保预测的动作轨迹得到更好的时序执行。

4.4. 自适应频率预测可视化

图8 展示了模型根据当前多模态观测所推断的未来力方差，调整执行频率的能力。尽管模型在推理期间无法获取未来数据，我们仍将未来力方差（图 8b）用作事后指标，以验证其预测结果。结果表明，模型能够准确预测即将发生的物理交互。如图8c 所示，它提高了更新频率

就在进入预计会发生接触、需要高精度控制的阶段之前。相反，在几乎没有或完全没有接触的稳定时期，它会将频率降低至 1× 。这种自适应行为凸显了系统能够根据预测的接触事件进行调整，从而在关键交互期间确保精确控制，而无需手动制定规则或依赖先验知识。

5. 结论

本文提出了 FAVLA，这是一种力自适应视觉-语言-动作框架，旨在将“慢速”语义模态（视觉、语言和历史力信息）与“快速”本体感知状态及力信号融合到单个端到端视觉-语言-动作模型中。同时，我们提出了力注入式快慢架构，采用大型视觉语言模型提取视觉-语言-力上下文，并预测未来的接触力变化。与此同时，我们提出了一种动作专家，通过使用力适配器将接触力信息注入多个层中，利用高频力测量值进行响应式动作调整。基于这一架构，我们进一步提出一种力自适应快慢推理策略，利用预测的力变化调整动作专家的执行频率。这样，模型便能及时修正动作轨迹，尤其是在接触力较大的情况下。在富接触工业任务上的大量真实机器人实验表明，FAVLA 的总体成功率达到 80.8%，并降低了峰值接触力；其成功率平均比基线视觉-语言-动作方法高 13.8%。这些结果表明，在端到端视觉-语言-动作模型中采用力自适应快慢策略，是实现反应式、高精度操作的有效途径。

影响声明

本文旨在推动机器人操作领域的深度学习发展，尤其是富接触、高精度工业任务方面的发展。通过在视觉-语言-动作模型中实现力自适应反应式控制，我们的方法有望提升物理环境中运行的机器人系统的安全性、可靠性和效率。这些能力可能有益于工业自动化及相关应用。

参考文献

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschan- nen, M., Bugliarello, E. 等。Paligemma：一种用于迁移的多功能 3B 视觉语言模型。*arXiv* 预印本 *arXiv:2407.07726*，2024。

Bi, J., Ma, K. Y., Hao, C., Shou, M. Z., 和 Soh, H. VLA-Touch：通过双层触觉反馈增强视觉-语言-动作模型。（arXiv:2507.17294），2025 年。doi：10.48550/arXiv.2507.17294。

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B. 等。π_0：一种用于通用机器人控制的视觉-语言-动作流模型。*arXiv* 预印本 *arXiv:2410.24164*，2024。

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Galliker, M. Y., 等。$\pi_{0.5}$：一种具有开放世界泛化能力的视觉-语言-动作模型。发表于第九届机器人学习年度会议，2025 年。

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., 等。Rt-1：用于大规模真实世界控制的机器人变换器。*ArXiv* 预印本 *arXiv:2212.06817*，2022 年。

Chen, H., Liu, J., Gu, C., Liu, Z., Zhang, R., Li, X., He, X., Guo, Y., Fu, C.-W., Zhang, S., 等。Fast-in-slow：一种将快速操作与慢速推理相结合的双系统基础模型。*ArXiv* 预印本 *arXiv:2506.01953*，2025a。

Chen, W.、Xue, H.、Wang, Y.、Zhou, F.、Lv, J.、Jin, Y.、Tang, S.、Wen, C. 和 Lu, C. ImplicitRDP：一种采用结构化慢快学习的端到端视觉-力扩散策略。（ArXiv:2512.10946），2025 年 b。doi：10.48550/arXiv.2512.10946。

Cheng, Z., Zhang, Y., Zhang, W., Li, H., Wang, K., Song, L., 和 Zhang, H. OmniVTLA：具备语义对齐触觉感知的视觉-触觉-语言-动作模型。*ArXiv* 预印本 *arXiv:2508.08706*，2025 年。

Feng, R., Hu, D., Ma, W., 和 Li, X. 按乐谱演奏：面向机器人操作的阶段引导的动态多感官融合。*ArXiv* 预印本 *arXiv:2408.01366*，2024。He, Z., Fang, H., Chen, J., Fang, H.-S., 和 Lu, C. Foar：面向接触丰富型机器人操作的力感知反应策略。*IEEE*机器人与自动化快报，2025 年。Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. 等。LoRA：大语言模型的低秩适配。*ICLR*，1(2):3，2022。

Huang, B.、Wang, Y.、Yang, X.、Luo, Y. 和 Li, Y. 3D-ViTac：利用视觉-触觉感知学习精细操作。（ArXiv:2410.24091），2025 年 a。doi：10.48550/arXiv.2410.24091。

Huang, J.、Wang, S.、Lin, F.、Hu, Y.、Wen, C. 和 Gao, Y. Tactile-VLA：释放视觉-语言-动作模型的物理知识，实现触觉泛化。（ArXiv:2507.09160），2025 年 b。doi：10.48550/arXiv.2507.09160。

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakr- ishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., San- keti, P., 等。OpenVLA：一种开源视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2406.09246*，2024。

Langley, P. 撰写机器学习论文。载于 Langley, P.（编），第 *17*届国际机器学习会议论文集（*ICML 2000*），第 1207–1216 页，加利福尼亚州斯坦福，2000 年。Morgan Kaufmann。

Li, J., Wu, T., Zhang, J., Chen, Z., Jin, H., Wu, M., Shen, Y., Yang, Y. 和 Dong, H. 用于灵巧操作的自适应视觉-触觉融合与预测性力注意力。（arXiv:2505.13982），2025 年。doi：10.48550/arXiv.2505.13982。

Liu, W., Wang, J., Wang, Y., Wang, W. 和 Lu, C. Forcemimic：使用力运动捕捉系统进行以力为中心的模仿学习，以实现富接触操作。载于*2025*年*IEEE*国际机器人与自动化会议 *(ICRA)*，第 1105–1112 页。IEEE，2025 年。

Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C. 和 Levine, S. Fast：面向视觉-语言-动作模型的高效动作标记化。*ArXiv* 预印本 *arXiv:2501.09747*，2025 年。

Shahidzadeh, A.-H., Caddeo, G. M., Alapati, K., Natale, L., Ferm¨uler, C. 和 Aloimonos, Y. Feelanyforce：利用触觉感知估计视觉式触觉传感器的接触力反馈。载于*2025* 年 *IEEE* 国际机器人与自动化会议 *(ICRA)*，第 251–257 页。IEEE，2025 年。

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E. 和 Dean, J. 超大型神经网络：稀疏门控混合专家层。*ArXiv*，abs/1701.06538，2017 年。网址：
h
ttps://api.semanticscholar.org/CorpusID:12462234。

Wen, J., Zhu, Y., Zhu, M., Tang, Z., Li, J., Zhou, Z., Liu, X., Shen, C., Peng, Y., 和 Feng, F. DiffusionVLA：通过统一扩散与自回归扩展机器人基础模型。在 第四十二届国际机器学习会议，2025 年。

Xue, H., Ren, J., Chen, W., Zhang, G., Fang, Y., Gu, G., Xu, H., 和 Lu, C. 反应式扩散策略：面向富接触操作的慢快视觉-触觉策略学习。（arXiv:2503.02881），2025 年。doi：10.48550/arXiv.2503. 02881。

Yu, J., Liu, H., Yu, Q., Ren, J., Hao, C., Ding, H., Huang, G., Huang, G., Song, Y., Cai, P., 等。ForceVLA：通过力感知混合专家增强视觉-语言-动作模型，以实现富接触操作。*arXiv* 预印本 *arXiv:2505.22159*，2025 年。

Zhai, A., Liu, B., Fang, B., Cai, C., Ma, E., Yin, E., Wang, H., Zhou, H., Wang, J., Shi, L., 等. 点燃视觉语言模型，迈向具身空间。*arXiv* 预印本 *arXiv:2509.11766*，2025 年。

Zhang, J., Guo, Y., Chen, X., Wang, Y.-J., Hu, Y., Shi, C., and Chen, J. HIRT：利用分层机器人变换器增强机器人控制。*arXiv* 预印本 *arXiv:2410.05273*，2024。

Zhang, J., Guo, Y., Hu, Y., Chen, X., Zhu, X., and Chen, J. UP-VLA：面向具身智能体的统一理解与预测模型。发表于第四十二届国际机器学习大会，2025a。

Zhang, Z., Xu, H., Yang, Z., Yue, C., Lin, Z., Gao, H.- a., Wang, Z., and Zhao, H. TA-VLA：阐释力矩感知型视觉-语言-动作模型的设计空间。*arXiv* 预印本 *arXiv:2509.07962*，2025b。

Zhao, T., Kumar, V., Levine, S. 和 Finn, C. 使用低成本硬件学习精细双手操作。*ArXiv*，abs/2304.13705，2023。网址：
h
t
tps://api.semanticscholar.org/CorpusID:258331658。

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A. 等。Rt-2：视觉-语言-动作模型将网络知识迁移至机器人控制。在机器人学习会议，第 2165–2183 页。PMLR，2023。

A. 附录

A.1. 遥操作系统

我们的遥操作系统如图9所示。操作员使用 3D SpaceMouse 在末端执行器空间中控制机器人。SpaceMouse 的相对运动会经过缩放，并映射为机器人的末端执行器相对运动。为弥补缺少力反馈的问题，我们将力读数转换为音频信号。6 个力维度分别映射到不同的音频频率（220–880 Hz），使操作员能够识别受力方向。音量与力的大小成正比。操作员佩戴耳机，通过实时音频反馈感知接触力的方向和大小。

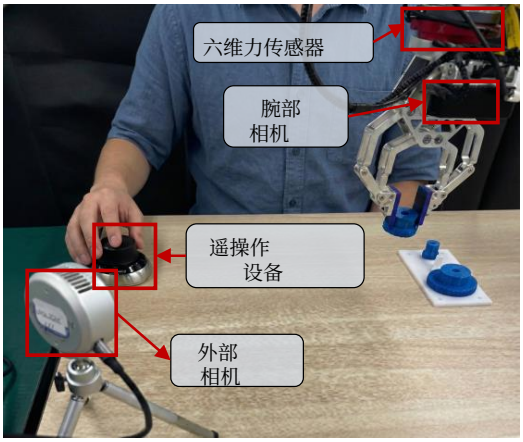


图 *9.* 遥操作系统设置。

A.2. 数据采集详情

我们通过人工遥操作收集演示数据，使用第A.1节所述的系统。操作人员通过 3D SpaceMouse 控制机械臂完成操作任务，同时系统同时记录 轨迹数据，包括机器人状态、力/力矩测量和相机观测 s.

为了提高轨迹多样性，我们会在每次演示开始时随机设置机械臂和任务物体的初始位置。我们发现，改变位置有助于模型学习到泛化能力更强的操作策略。

遥操作期间，我们以不同频率记录多路数据流：以 200 Hz 记录末端执行器位姿和六维力/力矩读数，并以 30 Hz、分辨率为 640×360 记录外部相机和腕部相机图像。所有数据流均经过同步，并下采样至 30 Hz，以实现时间对齐。随后，将收集到的数据集转换为 LeRobot v2.1 格式，用于后续训练。

A.3. 数据集统计

我们在表 3 中给出了所收集数据集的数据集统计，包括每项任务的轨迹数和帧数。我们的数据集包含四项富接触操作任务，共有 260 条轨迹和 198,250 帧，相当于约 1.84 小时的真实世界时间。尽管数据集规模相对较小，我们发现轨迹的高度多样性使模型能够有效地学习如何执行这些任务。

表 *3.* 四项操作任务的数据集统计。

Task	轨迹数	帧	小时数
USB 插入	80	44,716	0.41 小时
齿轮装配	80	58,038	0.54 小时
盒子翻转	50	47,476	0.44 小时
板面擦拭	50	48,020	0.45 小时
总计	**260**	**198,250**	**1.84 小时**

A.4. 任务描述

USB 插入。 此任务的目标是将 USB 闪存盘完全插入标准 USB 接口。机器人最初用夹爪夹持着 USB 闪存盘。要成功完成任务，它必须先移动到目标接口上方，将闪存盘与接口对齐，最后执行插入操作。此任务要求达到亚毫米级精度，因为即使轻微错位也可能导致插入失败。

齿轮装配。 目标是将机器人夹爪夹持的齿轮安装到底座上的中央立柱上，确保齿轮的齿与两侧预装的相邻齿轮完全啮合。这项任务面临两大挑战：首先，必须将齿轮的孔与立柱对齐。由于孔不可见，机械臂必须依靠实时力反馈来

表4。 观测与动作空间的规范。

字段	组成部分	形状
动作（块大小 = 32）		
末端执行器位姿增量 + 夹爪宽度	$\Delta x, \Delta y, \Delta z, \Delta\phi, \Delta\theta, \Delta\psi, g$	32×7
观测		
末端执行器位姿 + 夹爪宽度	$x, y, z, \phi, \theta, \psi, g$	7
力/力矩历史（10步）	$f_x, f_y, f_z, \tau_x, \tau_y, \tau_z$	10×6
头部图像	RGB	$360 \times 640 \times 3$
腕部图像	RGB	$360 \times 640 \times 3$
文本指令	分词后的文本序列	48

确定是否对齐。其次 nd，为了正确啮合，机器人还需要利用力反馈来确保齿轮对齐
在最后一次接触过程中

箱体翻转。 此任务旨在将水平放置的纸板箱翻转至垂直状态。机器人需要适应不断变化的接触力，并根据力反馈调整施加力的方向和大小，以避免用力过大损坏纸板箱，或因接触力不足导致纸板箱滑动。

板面擦拭。 此任务要求机器人抓起板擦，擦除白板上的标记。挑战在于确保板擦与白板表面稳定接触，从而施加足够的力以有效擦除标记。

A.5. 训练细节

观测与动作空间定义见表4。机器人的输入由多模态观测组成，包括来自腕部和外部相机的 RGB 图像、本体感觉状态、六轴力/力矩数据以及语言指令。输出为块大小为 32 的动作块。

我们在表 5中介绍模型架构和参数。我们采用 π_0 架构，该架构集成了视觉语言模型（VLM）和动作专家，并新增了一个用于预测力方差的力方差网络。视觉语言模型和动作专家使用 π_0 预训练权重进行初始化，并通过 LoRA 对模型进行微调；力方差网络则从头开始训练。

我们在每个数据集上训练一个任务特定模型，进行 30,000 次迭代，批量大小为 8，使用 AdamW优化器和余弦衰减学习率计划（初始学习率为 2.5×10^{-5}，最低学习率为 1×10^{-6}）。在单张 NVIDIA A100 80GB GPU 上，训练约需 6 小时。

为增强鲁棒性，我们采用数据增强：对非腕部图像进行随机裁剪和 $\pm 15°$ 旋转，并对所有摄像头的图像进行颜色调整，分别将亮度、对比度和饱和度调整 ± 0.2 。

A.6. 定性结果

真实世界可视化。 如图所示 10、图 11、图12和图 13所示，我们详细展示了各项任务的执行过程。观测结果表明，我们的模型能够适应各种扰动，包括角度旋转、未见过的颜色以及不同形状的箱体，并具有很强的泛化能力和鲁棒性。在 USB 插入任务中，模型能够适应经过角度旋转的插座，并通过力反馈调整自身位置，即使未对准也能实现精准插入。在齿轮装配任务中，模型对未见过的齿轮具有很强的泛化能力，包括训练中未出现过的颜色，并能够成功插入这些齿轮。在箱体翻转任务中，模型具备出色的力控制能力，可防止箱子滑动，并能有效适应各种未见过的箱子尺寸。在板面擦拭任务中，模型能够泛化到颜色和文字各异的标记。总体而言，我们的模型在真实世界操作任务中展现出很强的鲁棒性和泛化能力，验证了其实际适用性和可靠性。

失败案例分析。 我们对全部四项操作任务中的典型失败案例进行了系统分析，以找出关键局限及其根本原因，如图 14 所示。

表5. 模型架构和超参数概要

模块	配置
视觉-语言模型	
总参数量	2.6B
Transformer层	18
隐藏维度	2048
FFN隐藏维度	16384
注意力机制	分组查询注意力，8 个查询头、1 个键值头，头部维度 = 256
LoRA	秩 $r = 16$，缩放因子 $\alpha = 16$
动作专家	
总参数量	300M
Transformer层	18
隐藏维度	1024
FFN隐藏维度	4096
注意力机制	分组查询注意力，8 个查询头，1 个键值头，头部维度 = 256
LoRA	秩 $r = 32$，缩放因子 $\alpha = 32$
TCN力编码器	
总参数量	25.2M
TCN 块	4
隐藏维度	1024
卷积核大小	3
膨胀率	$[1, 2, 4, 8]$
Dropout 率	0.1
激活函数	Swish
输出力 token 数量	4
力方差预测器	
总参数量	1.6M
MLP 隐藏维度	（256，128）
输入维度	6144
输出维度	1
Dropout 率	0.1
激活函数	GELU

对于 USB 插入任务，机器人有时会在成功插入之前松开夹爪；这种情况仅发生在 USB 闪存盘与插槽错位时。此类失败是由于机器人错误理解了力反馈：错位产生的较大接触力与 USB 闪存盘完全插入时的力相似，导致机器人误以为任务已完成，便松开夹爪。

对于齿轮装配任务，主要故障是接触力过大并超出安全限值，导致系统停止工作。当齿轮与中央立柱未正确对齐时，就会出现这个问题；机器人仍然向下施力以插入齿轮，导致接触力过大并触发安全停止。

对于箱体翻转任务，主要故障是接触力不足，导致箱体滑动。出现这个问题是因为机器人为了避免损坏物体而施力过小，接触力不足进而导致箱体滑动。

对于板面擦拭任务，主要故障是白板未擦干净，白板上仍有明显的墨水残留。出现这个问题是因为擦拭轨迹不正确，机器人未能擦到上方的标记，导致白板未擦干净。

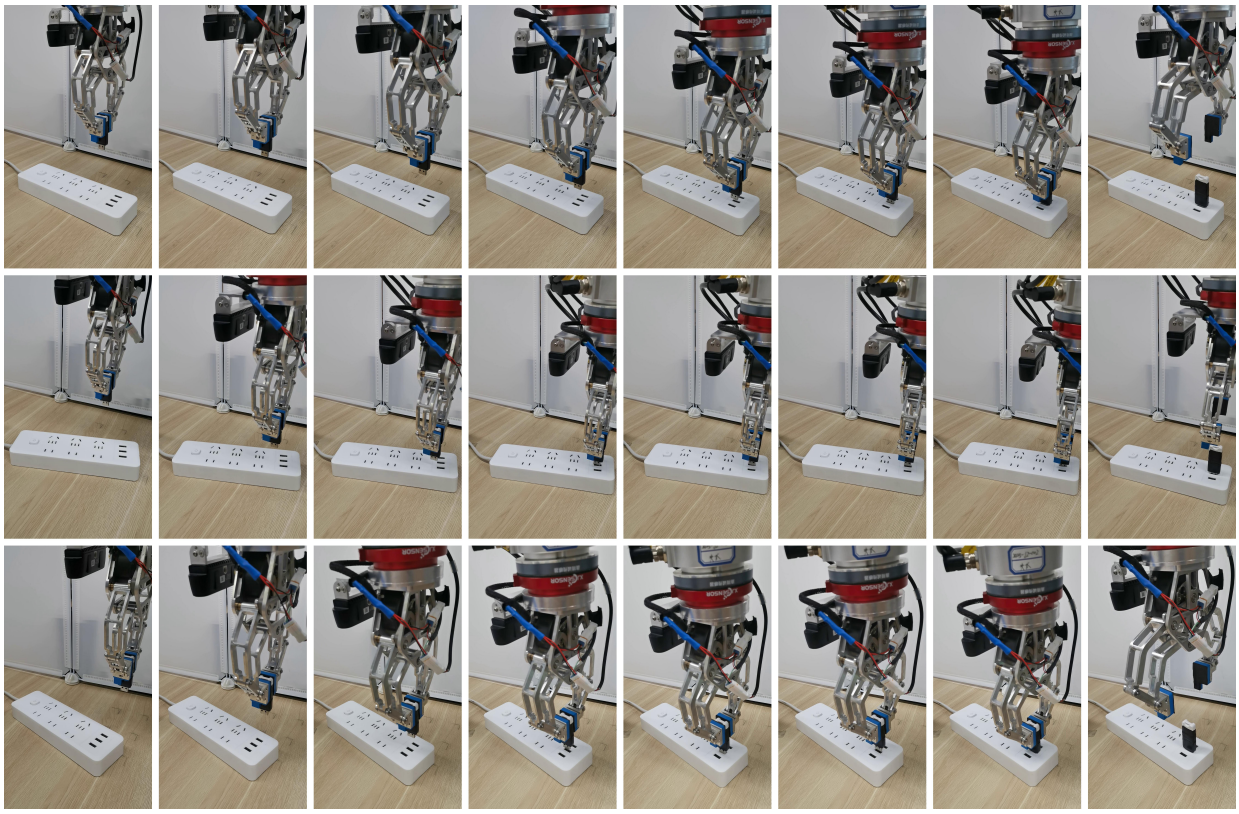


图 *10.* 真实世界**USB 插入** 任务的定性结果。我们的模型成功应对了高达 30° 的插线板旋转。

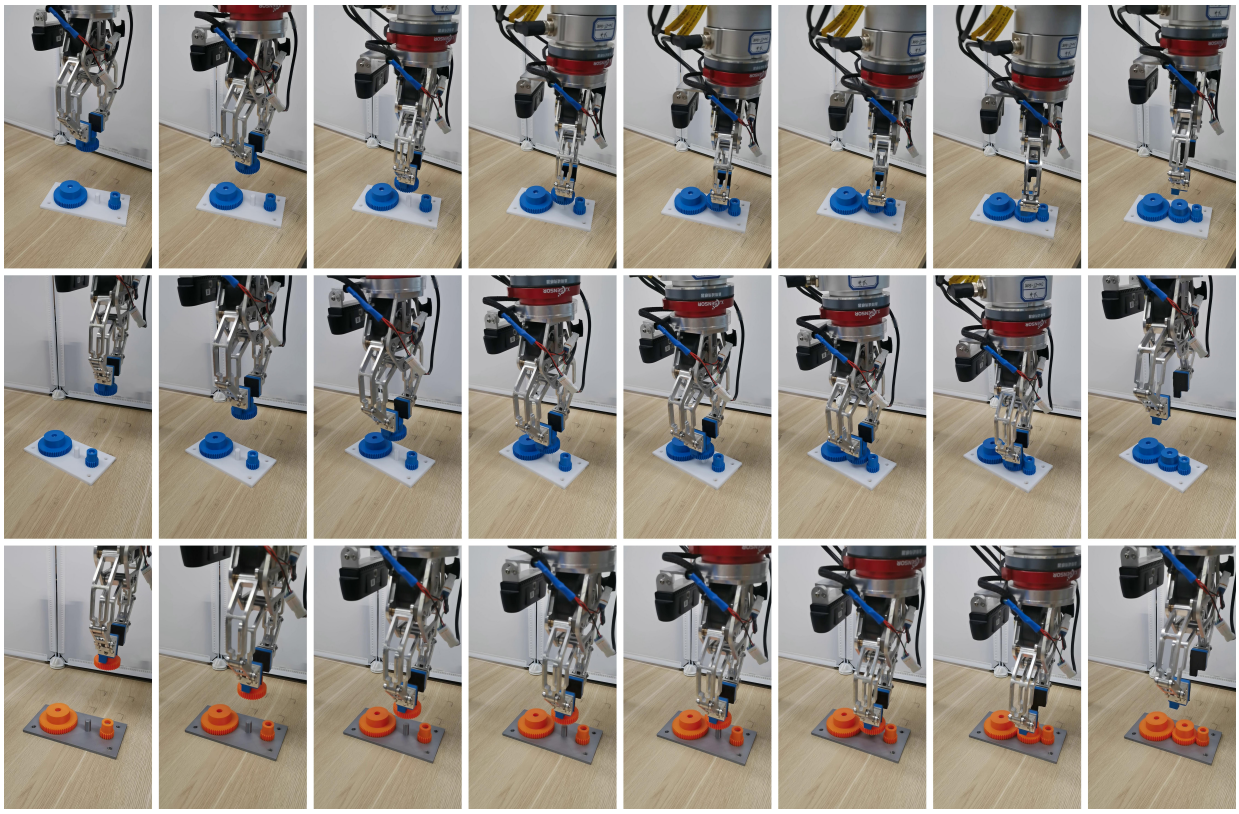


图 *11.* 真实世界**齿轮装配** 任务的定性结果。我们的模型成功泛化到未见过颜色的齿轮。

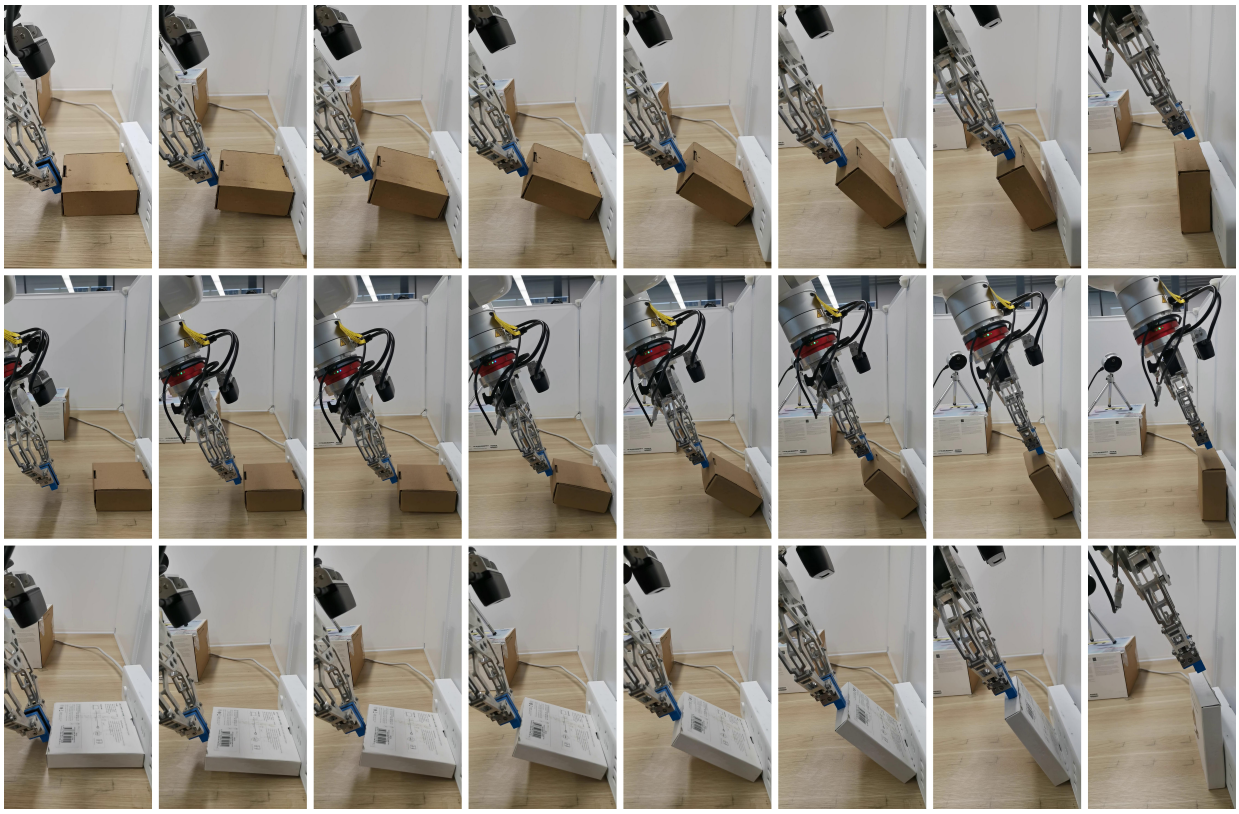


图*12.* 真实世界中**盒子翻转**任务的定性结果。我们的模型能够泛化到尺寸未见过的箱体。

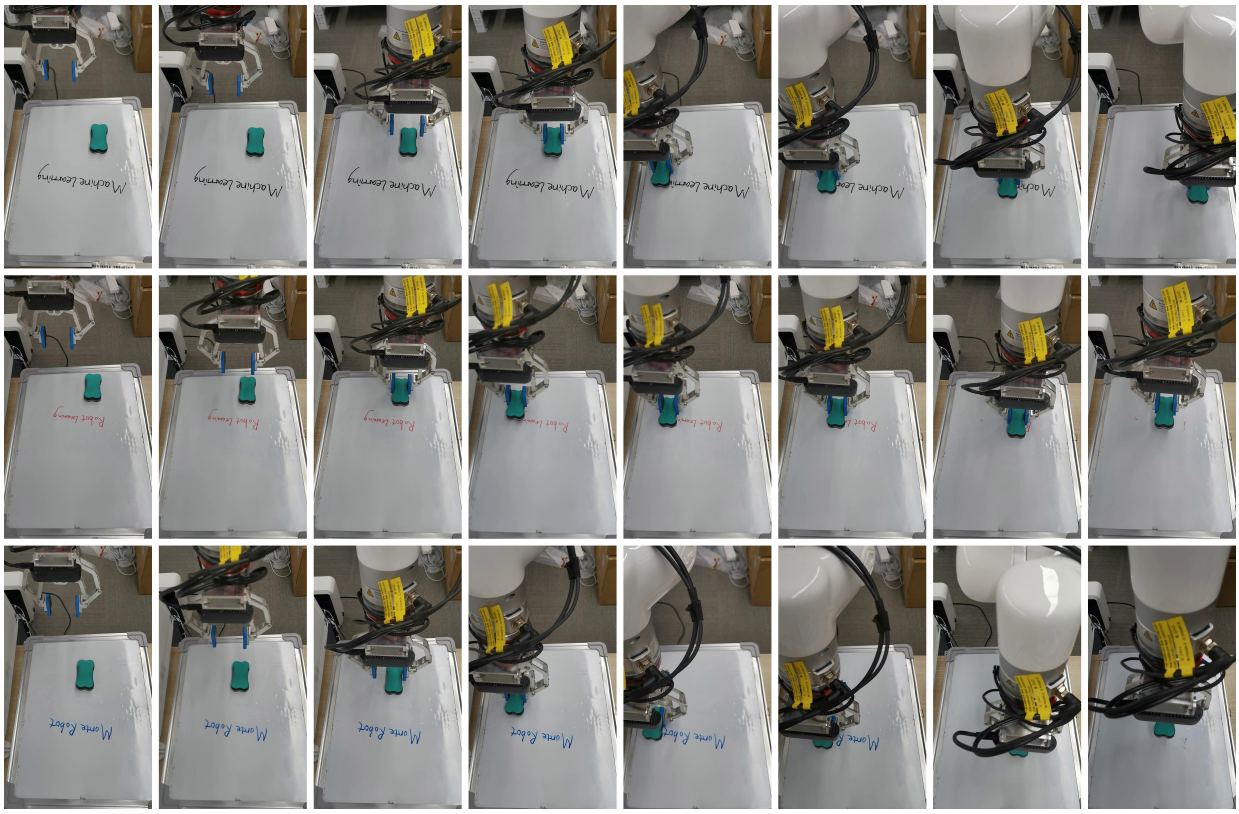
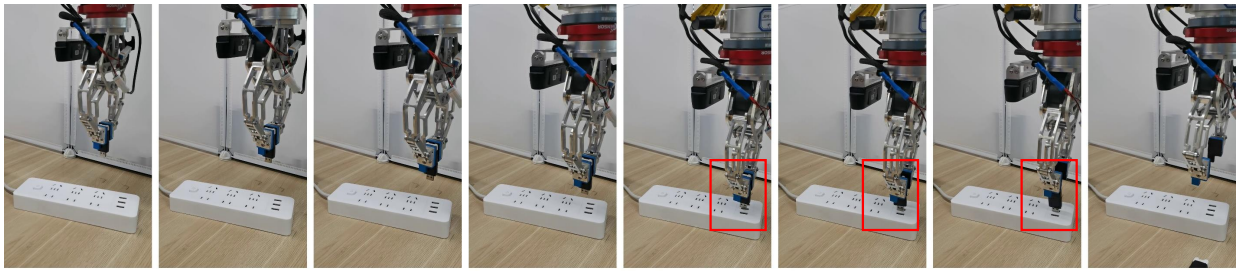
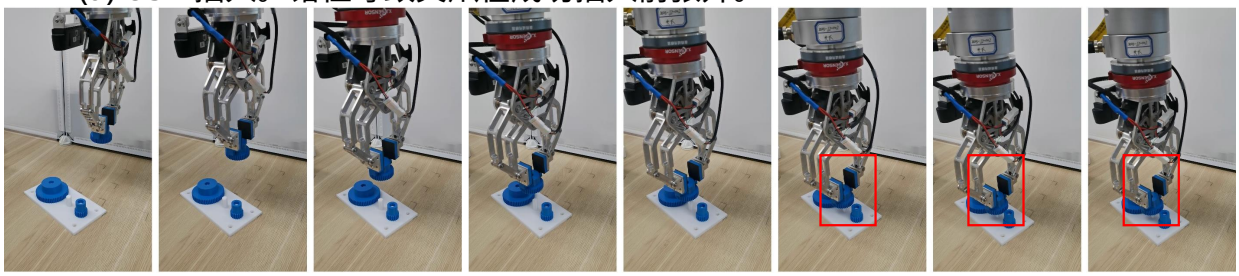


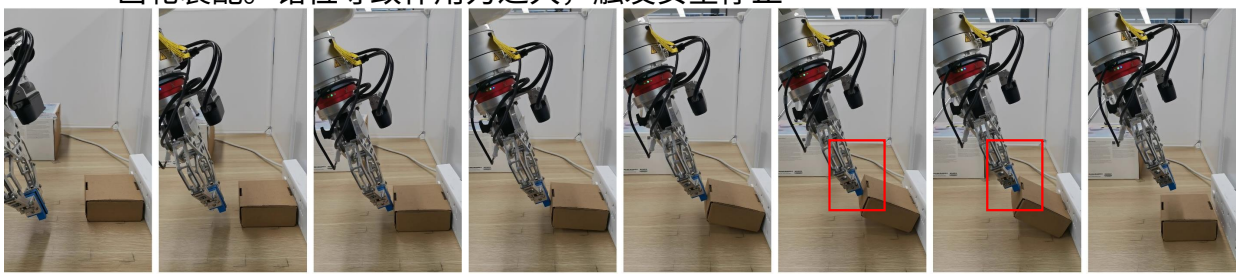
图*13.* 真实世界中**板面擦拭** 任务的定性结果。我们的模型能够泛化到颜色未见过的记号笔墨水。



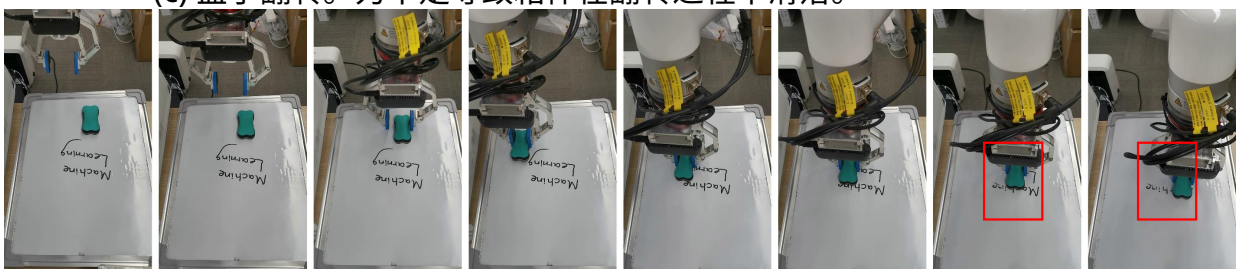
(a) USB 插入。错位导致夹爪在成功插入前张开。



(b) 齿轮装配。错位导致作用力过大，触发安全停止 .



(c) 盒子翻转。力不足导致箱体在翻转过程中滑落。



(d) 板面擦拭。错误的擦拭轨迹导致白板上留下墨水残留。

图 *14*。 不同任务中失败案例的可视化。