arXiv:2603.01700v1 [cs.RO] 2 Mar 2026

TacMamba：连接快速反射与慢速视觉-语言-动作推理的触觉历史压缩适配器

Zhenan Wang[1, 2, †], Yanzhe Wang[1, †], Meixuan Ren[1,3],Peng Li[2],Yang Liu[1],
Yifei Nie[1,2], Limin Long[2], Yun Ye[2],Xiaofeng Wang[2],Zhen Zhu[2] 和 Huixu Dong[1, *]

摘要—在视觉模糊的操作任务中，例如检测按钮的“咔嗒声”，触觉反馈往往是真实依据的唯一来源。然而，由于存在时空不匹配，融合触觉数据面临重大挑战：触觉感知需要高频处理和长时程记忆（系统1），而视觉策略则以低控制频率运行（系统2）。现有架构难以弥合这一差距：Transformer在高频循环（$>$ 100Hz）中计算开销过大，而长短期记忆网络则难以避免在长时间交互历史中遗忘。本文提出TacMamba，这是一种分层架构，可使高带宽触觉反射与低频视觉规划相协调。我们的方法包含三项核心贡献：（1）定制的高频触觉接口，旨在实现灵活集成；（2）基于Mamba的触觉历史压缩器，将连续力历史编码为紧凑状态，并以$\mathcal{O}(1)$推理延迟（0.45毫秒）运行，无需联合预训练即可与视觉-语言-动作模型即插即用地融合；（3）触觉引导的双阶段训练策略，利用时间判别进行自监督表征学习，并通过阶段均匀采样缓解数据稀疏性。在离散计数和隐式状态切换任务上的实验表明，TacMamba的成功率达到100%，显著优于纯视觉基线$\pi_{0.5}$，同时严格满足硬实时约束。

I. 引言

人类灵巧性将缓慢的语义规划（系统2）与快速的反射性反应（系统1）协同起来 [1]。尽管近期的视觉-语言-动作（VLA）模型已在系统2能力方面展现出熟练水平 [2], [3], [4], [5],[6], [7], ，但它们难以模拟系统1的反射特性。根本瓶颈在于频率不匹配：现有架构无法高效处理高频触觉历史，从而在高层视觉推理与低层物理交互之间造成脱节。

这种差异在以视觉歧义为特征的场景中至关重要，例如检测细微的按钮“咔嗒声”或识别滑牙螺纹 [8], [9], [10]。在此类情况下，触觉反馈是不可或缺的真实依据。与捕捉离散快照的视觉不同，触觉感知依赖于接触时间特征——连续的力曲线和瞬态振动。因此，稳健的操作需要具备长期时序记忆的模型，以便将成功触发与失败区分开来，其依据是交互历史，而非瞬时状态 [11], [12]。

然而，对这一延长历史进行建模会带来难以承受的计算负担。高分辨率传感器以超过100Hz的频率流式传输数据 [13], [14]。仅用标准Transformer处理两秒的历史就会生成数百个词元 [15],，导致计算成本呈二次方增长（$\mathcal{O}(N^2)$）。这种延迟成为高频控制回路的瓶颈，实际上将“快速”系统限制在“慢速”视觉规划器的响应时间内。相反，传统循环神经网络（例如长短期记忆网络）虽能实现快速推理，但会遭受灾难性遗忘，无法在数千个时间步上保留细粒度的物理细节 [16], [17], [18]。

除了架构上的局限，多模态训练还面临两个严重瓶颈。首先，通过端到端协同训练将高频触觉数据流整合进大型视觉-语言-动作模型，计算成本高得难以承受 [3], [2] ，还可能导致已建立的语义先验发生灾难性遗忘 [9], [8]。其次，密集接触任务存在极端的时间数据不平衡 [19]。关键交互事件，例如按钮的“突跳”（图1），持续时间极短（占轨迹的$<$ 5%） [20]。标准的均匀采样会让这些重要信号淹没在无关紧要的伸手动作中，导致策略恰恰在接触瞬间失效。

为解决这些时空与数据方面的挑战，我们提出**TacMamba**，一种统一的软硬件框架。如图1所示，TacMamba通过异步架构弥合频率差距：触觉编码器（系统1）以100Hz连续流式推理，更新压缩后的触觉历史隐藏向量；随后，低频（$\sim$1Hz）的视觉-语言-动作规划器（系统2）按需查询该向量。我们的贡献主要有三点：

(1) 我们提出 TacMamba，这是首个无需下采样、能够将高带宽触觉反射（100Hz）与低频视觉推理（$\sim$1Hz）对齐的分层架构。不同于会丢弃高频数据的以往方法，我们的设计能够无损、实时地处理微尺度接触动力学。

(2) 一种新颖的、基于 Mamba 的触觉编码器 [21], [22] ，解决了精度-延迟权衡问题。TacMamba 在不牺牲感知保真度的前提下，实现了常数时间 $\mathcal{O}(1)$ 推理，其速度显著优于 Transformer 基线。这是首个能够在嵌入式硬件的严格实时约束下处理超长触觉历史的架构。

（3）我们提出了一种稳健、即插即用的两阶段学习

†这些作者对本研究贡献相同。
*通讯作者。[1] 浙江大学，杭州，中国[2] GigaAI，北京，中国[3] 哈尔滨工业大学，哈尔滨，中国

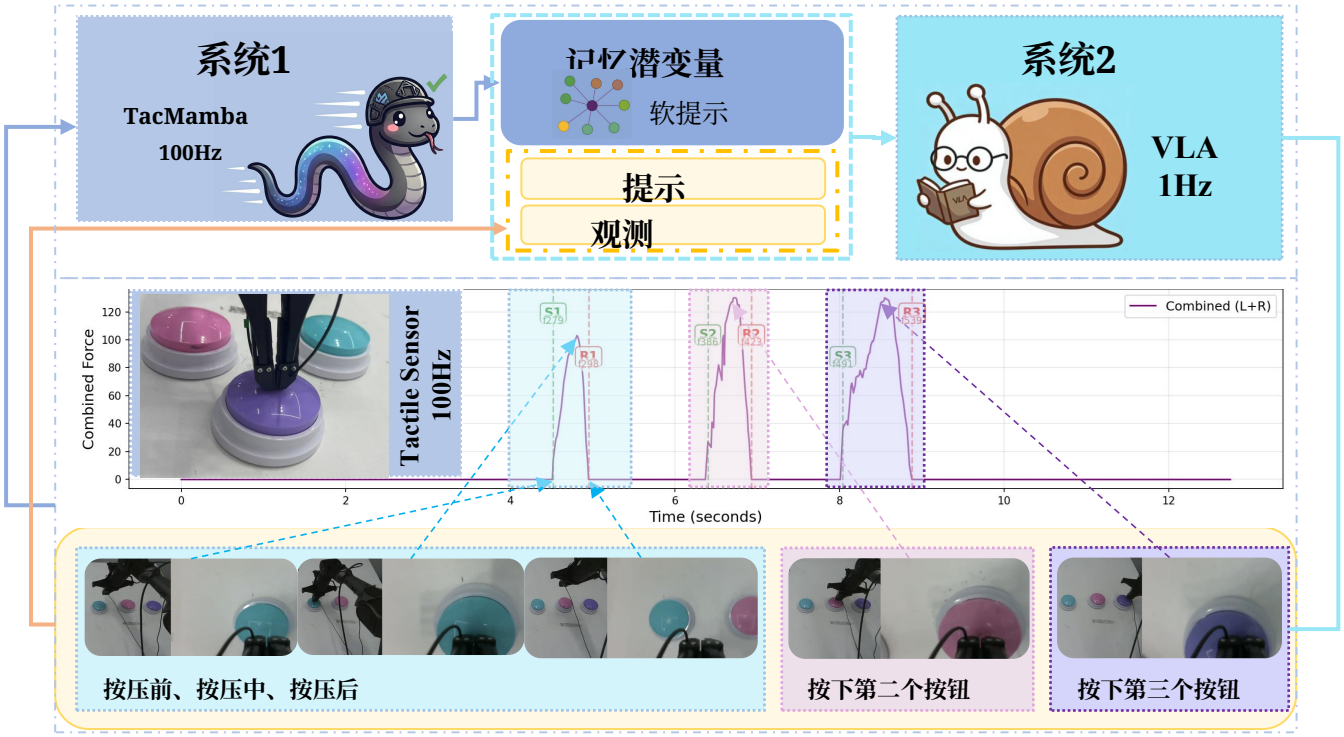


图1. **TacMamba系统架构**。该框架弥合了高速反射与低频推理之间的时空差异。**左侧（系统1）：** 触觉编码器使用Mamba模型以100Hz处理一维力信号流，并实时递归更新隐藏状态 h_t 。**右侧（系统2）：** 该压缩后的隐藏状态 h_t 经过投影后，以软提示的形式异步注入低频（∼1Hz）视觉-语言-动作（VLA）规划器。

框架，无需进行成本高昂的视觉-语言-动作联合预训练。该框架包括：（i）通过三元时序判别进行自监督预训练，以独立捕捉因果接触动力学；（ii）采用新颖的阶段均匀采样策略增强视觉-语言-动作微调，从而实现高效微调。这种方法解决了关键交互帧稀疏的问题，并将触觉历史无缝融入视觉-语言-动作模型，显著提升数据效率。

II. 相关工作

A. VLA中的时间模型

视觉-语言-动作模型近期的进展展现出显著成效，使机器人具备了高层语义推理能力（系统2） [3], [4], [2]。在基础模型的基础上，近期研究进一步增强了视觉-语言-动作模型在开放世界场景中的泛化能力 [5], [6]。此外，为应对长时程任务中的状态混叠挑战，近期的框架集成了记忆机制。例如，MAP-VLA [12] 通过检索增强提示，利用历史演示数据构建记忆库；而MemoryVLA [11] 则维护感知-认知记忆。然而，这些增强记忆的视觉-语言-动作模型存在一个关键局限：它们只能以视觉规划器较低的控制频率（∼1-10Hz）运行。由于其记忆库完全依赖低频视觉或本体感觉状态，因此天然缺乏所需的高带宽物理反馈以进行实时反射调节，始终受限于“慢速”的系统2范式。

B. 触觉增强机器人操作

鉴于纯视觉模型在密集接触任务中的局限性，近期研究尝试将触觉和力模态融入VLA架构。ForceVLA [9] 引入了一个具备力感知能力的专家混合模块，在动作解码过程中整合6自由度力反馈；而OmniVTLA [8] 则采用语义对齐触觉编码器来处理多模态感官输入。尽管取得了这些进展，现有的触觉视觉-语言-动作模型仍受同步架构的根本性瓶颈所限。由于它们将触觉或力令牌直接输入庞大的VLA骨干网络，感官处理速度严格受限于VLA较低的推理速度。因此，这些模型只能以大幅降低的频率利用瞬时触觉快照，无法维持连续、高频的历史记忆。这种有损整合会丢弃关键的高频瞬态（例如突然的微滑移或按钮的突跳振动）。相比之下，TacMamba将感官反射与视觉规划完全解耦，无需受低频VLA的瓶颈限制，即可连续、无损地编码100Hz触觉历史。

C. 高效长序列建模

对大量历史序列进行建模是深度学习面临的一项根本挑战。尽管 Transformer 已被

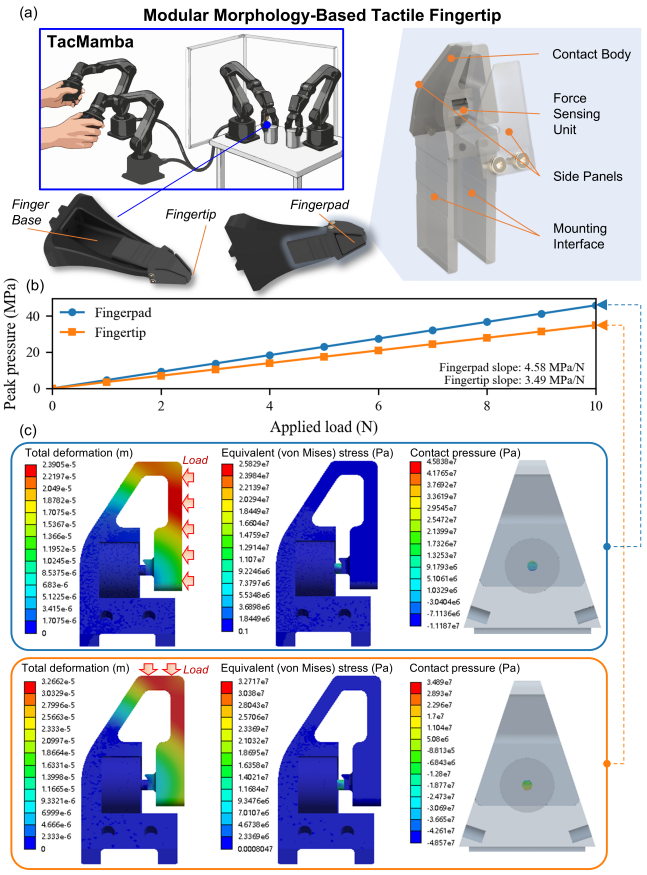


图 2. **系统概览**。(a) 模块化形态学触觉指尖设计，其中集成式柔顺接触体既支持指尖交互，也支持指腹交互，并通过机械结构将分布式接触投射到单轴力传感器上；侧面板保护内部结构，可重构双夹具接口则可将其安装到通用平行夹爪上。(b) 基于有限元法的力投射表征，展示指尖和指腹接触时峰值接触压力随施加载荷的变化；(c) 总变形、等效（冯·米塞斯）应力和接触压力的代表性可视化结果。

应用于长期时间序列预测， [23], [24],但其$\mathcal{O}(N^2)$延迟使其无法用于高频嵌入式控制回路。相比之下，传统的循环神经网络（RNN）推理速度快，却出了名地容易在数千个时间步中遭遇灾难性遗忘 [16], [17]。近年来，结构化状态空间模型（SSM） [22] 及其选择性变体 Mamba [21],作为一种范式转变崭露头角。Mamba 通过维持压缩的隐藏状态，实现了线性时间序列建模和严格的 $\mathcal{O}(1)$推理延迟 [25]。受此启发，TacMamba 开创了将 Mamba 应用于高频触觉流的先河，有效解决了记忆与延迟之间的权衡，并实现了系统1与系统2之间无损的频率对齐。

III. 硬件系统

TacMamba 是一个统一的软硬件系统，能够让高频触觉反射为低频视觉-语言-动作推理提供信息，而无需降采样。硬件系统概览如（图 2）所示。我们开发了一种模块化形态学触觉指尖，利用集成式柔顺接触体，依据仿生软传动原理，将分布式表面接触机械地映射到单一感测轴上 [26]。该模块由柔顺接触体、单轴力感测单元、侧向面板，以及可重构的双夹具安装接口组成；该接口可连接通用平行夹爪，并可适配不同的夹爪几何形状（图 2a）。与将远端感测和掌侧感测分开的设计不同，指尖和指腹区域采用单一连续柔顺结构实现，使指尖交互（例如按按钮）和指腹交互（例如表面接触和包覆）都能通过连续的力传递路径传导，并被机械地汇集至力传感器，形成统一的一维触觉流。侧向面板可遮蔽内部传动结构、提供机械保护，并在接触过程中限制不必要的侧向变形 [27]。

为量化空间位置不同的接触如何映射到传感器接口，我们在 0 至 10N 的施加载荷下进行有限元分析，并报告传感器表面上的总变形、等效（冯·米塞斯）应力和接触压力（图 2c）。所得压力-载荷曲线表明，指尖和指腹接触均呈现近似线性映射，其中指腹配置表现出更高的压力增益（图 2b）。这些结果证实，通过柔顺结构，不同表面区域施加的接触能够稳定地投射到单一感测轴上，从而形成可预测且单调的力传递路径。

我们强调，该设计并非旨在仅凭触觉感知恢复明确接触位置。相反，我们有意采用单轴力传感器作为系统级权衡。在我们的框架中，粗略接触位置已可部分从VLA骨干网络提供的视觉观测中获得，而按压贯穿、突跳和释放等关键交互线索主要表现为视觉难以捕捉的高频时序力变化。因此，一维力信号足以通过提供精确的幅值与动态反馈来补充视觉感知，这与聚合力历史包含丰富复杂操作信息的研究发现一致 [28]。这种极简触觉接口与TacMamba将快速物理反射注入缓慢视觉推理的目标相一致。

IV. 方法

TacMamba弥合了高速触觉反射（系统1）与低频视觉规划（系统2）之间的频率差距。所提出的框架包含两个核心组件：（A）面向连续时间触觉建模的硬件感知Mamba编码器，以及（B）通过自适应采样强制时间因果性的双阶段训练策略。

A. 基于*Mamba*架构的触觉编码器

我们采用 **Mamba**架构作为骨干网络，利用其核心选择性状态空间模型（SSM）机制。

这使我们能够对连续接触物理进行建模，同时通过依赖输入的选择性来处理突发的动力学不连续性（例如冲击）。

1）连续-离散建模：我们将一维触觉信号 $x(t) \in \mathbb{R}$ 视为驱动潜在物理系统的连续时间强迫函数。该系统通过线性时不变常微分方程，将输入映射为潜在状态 $h(t) \in \mathbb{R}^N$：

$$\dot{h}(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t) \tag{1}$$

为与机器人硬件的离散控制周期（时间步长 Δt）对齐，我们使用零阶保持（ZOH）原理对公式 1 进行离散化，同时保持物理一致性：

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t \tag{2}$$

其中，$\bar{\mathbf{A}} = \exp(\Delta t\mathbf{A})$，而 $\mathbf{B}$ 为离散化系统参数。

2）通过选择性捕获混合动力学：标准线性时不变系统依赖固定参数，无法满足具有混合动力学特征的操作任务需求。Mamba 通过其核心选择机制解决这一问题，该机制将系统动力学参数化为输入的函数：具体而言，我们在离散化之前计算依赖于输入的参数（$\Delta_t, \mathbf{B}_t, \mathbf{C}_t$）$= \text{Linear}(x_t)$。这样，模型便能作为自适应物理滤波器运行，在静态抓取时抑制噪声，同时在检测到“咔嗒声”等瞬态接触事件时迅速更新其内部状态。

3) 高效的分层架构与 $\mathcal{O}(1)$ 推理：如图 3（上）所示，我们的网络处理多变量非平稳力信号$\mathbf{X}$。为应对传感器异质性，我们简要地应用 RevIN 和通道独立性来提取通用物理特征。核心骨干网络由堆叠的 Mamba模块组成（详见图 3 左下），可分层捕获从高频脉冲到长期稳定性趋势的特征。

该架构最关键的优势是具备常数时间推理能力。利用公式 2 中的递归形式，我们将压缩后的隐藏状态 h_{t-1} 保存在内存中，使新力读数 x_t 的更新严格为 $\mathcal{O}(1)$。这与 Transformer 形成鲜明对比，确保我们的“快速”触觉系统（100Hz）完全不受“慢速”视觉策略延迟的影响。

B. 训练策略

为缓解标注触觉数据稀缺以及关键接触事件固有的时间稀疏性问题，我们提出了一种稳健的两阶段方案。

1）阶段 *1*：通过时间判别：为了使 Mamba编码器能够在没有人工标签的情况下理解接触动力学，我们设计了一项三元时序判别任务。

自动分割：首先，我们将触觉流视为一系列物理子任务。我们根据

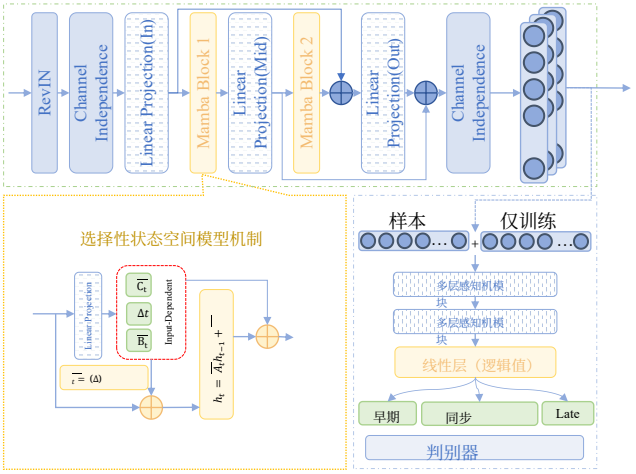


图3. TacMamba网络架构。上方：核心TacMamba主干网络通过分层选择性状态空间模型处理连续触觉流，并利用RevIN和通道独立性进行特征提取。左下方：Mamba模块机制的展开视图，说明输入相关参数（$\Delta t, B_t, C_t$）如何建模混合动力学。右下方：辅助判别器网络，仅在训练阶段使用，以促进稳健的特征学习并强制时间因果性。

触觉梯度幅值 $\|\nabla x_t\|$ 自动将长轨迹分割为离散的 子任务阶段，具体来说，就是检测从非接触到接触的转换。

目标：基于这些阶段，我们生成状态对（h_i, h_j），并训练判别器 D，将其时间关系分为三类：$y_{ij} \in \{0 : t_i < t_j, 1: t_i \approx t_j, 2: t_i > t_j\}$。关键在于，我们采用关系均衡采样策略，以严格保证这三个类别（早期、同时、晚期）上的均匀分布。这可防止模型退化为平凡的猜测策略（例如，总是预测“未来”），并迫使编码器将物理交互的因果结构嵌入隐状态空间。

我们通过最小化标准交叉熵损失来训练判别器：

$$\mathcal{L}_{pre} = -\mathbb{E}_{(h_i,h_j)\sim\mathcal{D}_{bal}} \sum_{c=0}^{2} \mathbb{I}(y_{ij} = c) \log D(h_i, h_j)_c \tag{3}$$

2) 阶段 *2*：采用阶段采样进行触觉引导*VLA*调优：在此阶段，我们冻结预训练的 Mamba 编码器，并将其压缩后的状态 h_t 作为“软提示”z_{tac} 输入视觉-语言-动作模型。

阶段均匀采样：机器人操作中的一项关键挑战是，与持续时间较长的空闲模式相比，关键交互帧（例如接触过渡）极为稀疏。如图 1 中稀疏的力峰值所示，关键事件仅占整个时间线的一小部分。

为此，我们利用阶段 1 的自动分割，将每条轨迹分解为一系列K离散子任务阶段 $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$。标准均匀采样（$p(t) \propto 1/T$）会使模型偏向持续时间较长的阶段（例如静态保持）。相比之下，我们采用阶段均匀采样策略，强制实现

各阶段之间的均匀分布：

$$p(t \in P_k) \propto \frac{1}{K \cdot |P_k|} \tag{4}$$

这确保每种子任务模式对训练批次的贡献相同，不受其持续时间影响。因此，短暂但关键的阶段（例如突跳发生的瞬间）与稳定模式具有同等权重，确保视觉-语言-动作模型学会完整的操作逻辑，而不会忽略短时触觉线索。

V. 实验

A. 实验设置

硬件平台：我们使用一台配备平行夹爪的 AgileX PiPER 轻型 6自由度机械臂。为捕捉接触动力学，我们在指尖集成了一个定制的单点高频力传感器，以 100 Hz 的频率传输一维力数据。本节旨在回答两个核心研究问题：

(1) 效率与可扩展性：TacMamba 能否在建模超长触觉历史的同时，维持高频闭环控制所需的实时响应能力？

(2) 模糊条件下的鲁棒性：在存在视觉遮挡或状态模糊性的场景中，注入高频触觉历史数据是否能显著提升操作成功率？

基线：我们将我们的方法与两类基线进行比较：

1) 骨干网络基线： 对于触觉编码器，我们将 TacMamba 与一维卷积神经网络（局部卷积）、长短期记忆网络（循环网络）和 Transformer（全局注意力）进行比较，以评估其历史信息压缩能力。2) 系统基线：对于端到端策略，我们将其与 $\pi_{0.5}$ [2] 进行比较。$\pi_{0.5}$ [2] 是一种最先进的通用视觉-语言-动作（VLA）模型，可用于评估专用触觉记忆的价值。

B. 性能评估：准确率与效率

我们在完全相同的数据集上，将 TacMamba 与标准架构进行基准测试。定量结果汇总于图 4 和表 I。

1) 综合分析：准确率与效率之间的权衡： 我们在增强的真实世界数据集上，对感知保真度与实时可行性之间的权衡进行全面评估。比较结果揭示了各类架构的不同特点：

TacMamba（本文方法）：TacMamba 达到 **最先进精度（88.89%）**，推理延迟极低，仅为 0.45 ms。其 $\mathcal{O}(1)$ 复杂度使其相较于长短期记忆网络实现了 21× 倍加速，且在不牺牲精度的情况下，独特地满足了高频控制要求（>100Hz）。

循环神经网络（长短期记忆网络）：双向长短期记忆网络的准确率为 82.50%，但每一步都需要重新评估完整历史记录，导致延迟达到 9.20 ms，违反了严格的流式处理约束。单向长短期记忆网络速度很快（0.54 ms），但在高频序列上会遭受灾难性遗忘，导致准确率降至 65.33%。

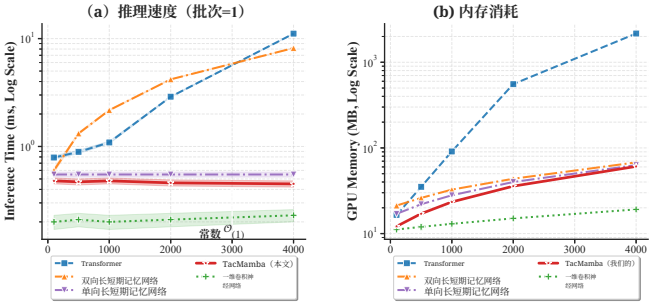


图 4. **效率分析**。推理延迟和内存增长。

表 I 增强的真实世界数据集上的比较

模型	Acc.	延迟（毫秒）	内存（MB）
一维卷积神经网络	56.25%	0.23 ± 0.02	19.14
Transformer	81.25%	11.37 ± 0.17	2164.71
双向长短期记忆网络	82.50%	9.20 ± 0.13	67.79
单向长短期记忆网络	65.33%	0.54 ± 0.01	63.70
TacMamba	**88.89%**	0.45 ± 0.02	60.85

Transformer：由于从含噪的触觉流中提取特征时数据利用效率低下，Transformer 的准确率仅为 81.25%。此外，其 $\mathcal{O}(L^2)$ 复杂度使延迟飙升至 11.37 ms，并导致内存使用量过高，因此无法用于边缘计算。

一维卷积神经网络：虽然计算效率较高（0.23 ms），但其准确率最低（56.25%）。其有限的感受野无法结合更广泛的任务历史来理解局部振动，从而导致状态混淆。

C. 真实世界验证：视觉遮挡下的连续交互

为全面验证在真实环境中的性能，并明确评估传统视觉模型在长时域场景中的局限性，我们设计了三项不同的操作任务——“连续按压按钮”“盲操作薯条包装”和“可变形衣物折叠”。这些任务旨在严格评测 TacMamba 与纯视觉基线 $\pi_{0.5}$ 在不同程度的视觉遮挡和状态模糊性下的表现。

基线选择策略：尽管我们在离线数据集上对长短期记忆网络和 Transformer 架构进行了广泛比较（第 V-B 节），但在这些真实世界的闭环任务中，我们不采用它们。系统1反射响应的一项关键限制是 100Hz（10 ms）的控制周期。然而，Transformer（复杂度为二次）和长短期记忆网络（顺序处理）的推理延迟都会随输入历史长度增加而增长。随着长时域操作过程中触觉历史不断增长，它们的计算时间不可避免地会超出这一严格的实时限制，造成危险的延迟。因此，TacMamba 凭借其恒定时间的 $\mathcal{O}(1)$ 推理，是唯一能够维持高频控制回路的架构，这使得 $\pi_{0.5}$ （仅视觉）成为唯一可行的系统级基线。

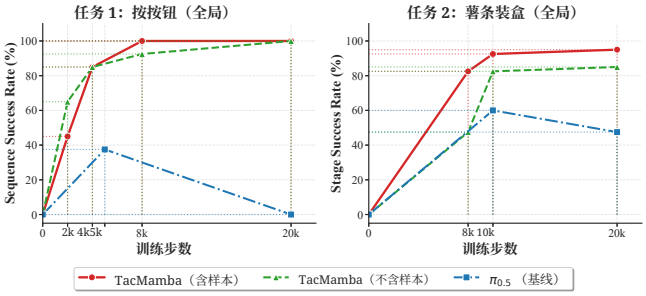


图 5. **全局任务成功率**。比较训练步数变化过程中的完整任务完成率。TacMamba（红色）能够快速收敛且鲁棒性高，而 $\pi_{0.5}$（蓝色）由于静态帧过拟合，在按钮任务中遭遇灾难性失败。

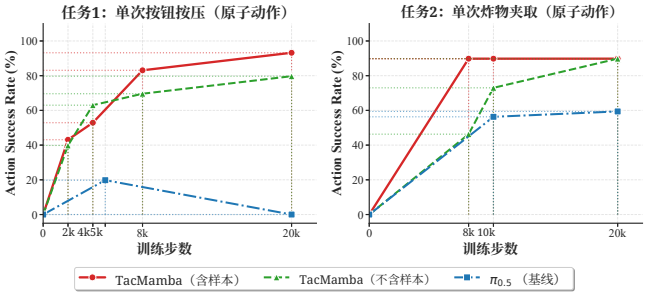


图 6. **原子动作成功率**。评估细粒度交互质量（例如，成功按下一次按钮且没有过冲）。TacMamba 能保持较高的物理精度，而基线则难以应对幻觉和执行错误。

1）任务定义：任务 1：连续按压按钮。机器人需要按照特定的颜色顺序按下三个按钮（蓝色 → 粉色 →紫色）。此任务带来了三项关键挑战：

1）视觉遮挡：操作过程中，机器人的末端执行器完全遮挡目标区域，导致$\pi_{0.5}$ 视觉编码器无法观察。2）状态模糊性：按钮触发状态（“咔嗒声”）仅提供细微的振动反馈，视觉上无法察觉。3）感知混叠（记忆缺口）：由于按钮外观完全相同，且视觉画面因遮挡而保持静止，仅依赖当前观察的模型无法区分操作序列所处的阶段（例如，第一次按压还是第二次按压）。要取得成功，必须持久记忆先前的交互。

任务2：盲操作薯条包装。机器人必须依次从盘子中取出三根薯条，并将它们放入不透明盒子中。其核心难点在于累积状态估计：

由于盒子内部被遮挡，外部视觉无法穿透，机器人实际上无法“看见”任务进度。系统必须完全依赖长时程触觉历史来跟踪计数（即记住此前成功取出的薯条数量），以确定何时终止任务。

任务3：可变形衣物折叠。如图7所示，机器人执行一个包含五个步骤的工作流程：将衣物分别向上、下、右、左四个方向折叠，然后放入篮子中。这一长程可变形物体操作严格考验视觉-语言-动作模型应对三项核心挑战的能力：

1）精确控制：织物极薄，因此需要进行精细、连续的物理交互。2）视觉歧义：动态褶皱很容易误导纯视觉模型，因此必须依靠触觉反馈来确认是否已稳固“捏取”。3）长时程过拟合：冗长的操作序列会加剧状态遗忘，并使策略过拟合于静态非接触画面，导致执行质量随时间下降。

评估指标：为细致分析物理交互质量，我们从两个层面报告性能：

1）全局任务成功率：机器人完成整个长时程序列的回合所占百分比（例如，按正确顺序按下全部 3 个按钮，或包装 3 根薯条）。

2) 原子动作成功率：采用严格标准，排除“侥幸”的子任务成功。按按钮：必须触发触觉“咔嗒声”并立即停止（过冲或虚按均视为失败）。薯条包装：必须牢牢抓取薯条、将其运送并释放到盒子里（抓空、掉落或边缘挂碰均视为失败）。

2) 结果比较与分析：我们在不同训练阶段评估 Tac-Mamba（使用和不使用我们的采样策略），并与 $\pi_{0.5}$ 进行比较。

为确保统计显著性并消除随机性，我们采用了严格的评估方案：对于每个训练检查点，我们在 10 个随机初始位置上进行实验，每个位置进行 5 次独立试验（每个数据点共计 50 个评估回合）。

表 II 顺序按钮按压任务的性能

方法	成功率
外部视觉基线*:	
OpenVLA [29]	6%
CogACT [30]	15%
MemoryVLA [11]	58%
我们的实现:	
$\pi_{0.5}$ [2](早期阶段，5k 步)	37.5%
$\pi_{0.5}$ [2]（已收敛，20k 步）	0%
TacMamba（阶段均匀，本文方法，20k 步）	**100%**

*外部视觉基线的结果直接来源于 [11]。其实验设置与我们完全相同，唯一的区别是使用了 Franka Panda机械臂，而不是我们的 AgileX PiPER。

1.基线局限性与对静态帧的时间过拟合：纯视觉基线（$\pi_{0.5}$）在长时程任务中表现吃力，如图 5 所示。在连续按钮按压任务中，观察到性能随时间退化：基线在 5k 步时的成功率为 37.5%，但到 20k 步时降至 0%。这种性能退化暴露了纯视觉VLA模型容易受到数据采样偏差导致的时间过拟合影响。由于标准随机采样会过度抽取静态非接触帧（这类帧在长时程序列中占主导），视觉



图 7. **折叠结果定性比较**。(a) 完整 TacMamba 模型执行的布料折叠流程。放大视图突显了高频触觉反馈带来的织物角部精准、稳定的抓取。(b) 和 (c) 为禁用阶段优先采样策略（消融实验）时观察到的典型失败模式：“抓取失败”（因视觉歧义而未能抓住角部）和“形状皱缩”（张力控制不当导致变形）。

策略会随着时间推移对这些静止特征过拟合。因此，收敛后的模型往往会悬停在按钮上方，失去执行动态按压动作的能力。这种性能退化并非个例；如表 II 所示，MemoryVLA [11]等视觉记忆架构在类似任务上的成功率停滞在 58%。这种性能上限源于视觉观测的低频特性：微尺度物理事件往往落在摄像头帧之间的盲区中，导致状态混淆。相比之下，TacMamba克服了对静态帧的过拟合，在收敛后仍保持 100% 的成功率，其策略以高频触觉反射为基础。同样，在薯条装袋任务中，视觉遮挡使 $\pi_{0.5}$ 的全局成功率仅为 47.5%。

2. TacMamba 的高频鲁棒性：TacMamba 有效绕过了这些感知瓶颈。如全局成功率（图 5）和原子动作成功率（图 6）曲线所示，我们的完整模型保持了出色的稳定性。通过以 100Hz 处理一维触觉流，它能够可靠地捕捉瞬态事件，并保持严格的物理精度。TacMamba 在 8k 步时按钮任务成功率达到 100%，在薯条任务中达到 95%，并借助长时域触觉历史，在没有视觉确认的情况下成功实现隐式计数。

3. 阶段均匀采样消融实验：将采用我们的阶段均匀采样训练的 Tac-Mamba 与标准均匀变体进行比较，凸显了该策略的数据效率和精细控制能力。如图 6 中明确跟踪的原子动作成功率所示，阶段均匀采样方法显著减少了微观执行误差（例如过冲或虚假抓取）。因此，在薯条任务中，训练 8k 步后，该方法的全局成功率达到 82.5%，接近均匀采样基线（47.5%）的两倍。通过明确优先处理稀疏、高频的接触事件，我们的策略将整体样本效率提升了约 2.5×。

4.可变形物体操作与硬件约束：由于织物极薄，尚未收敛的策略往往会施加过大的向下作用力，导致定制触觉传感器抵住刚性桌面时面临物理损伤风险。因此，我们仅在完全收敛（20k 步）后评估了衣物折叠任务。尽管所有模型的宏观折叠成功率都达到 100%，但细粒度质量各有差异。由于视觉皱褶难以辨识，视觉基线和均匀采样消融都出现了微观失败（例如“抓取失败”或“形状皱缩”）（图7）。完整 TacMamba 模型则保持了稳健的反射，在感知到织物边缘时能够可靠地恰好进行捏取。

六、结论

在本研究中，我们提出了 TacMamba，这是一种软硬件架构，成功弥合了高频触觉反射之间的关键时空差距

与低频视觉推理之间的关键差距。通过基于 Mamba 的架构对连续一维触觉流进行建模，我们的方法以严格的 $\mathcal{O}(1)$推理延迟（0.45 ms）达到了最先进精度（88.89%）。其性能以 25$\times$ 倍加速超越 Transformer 基线，并完全绕过其延迟瓶颈，成为首个能够在嵌入式硬件的硬实时约束内处理超长触觉历史（$L = 5000$）的架构。此外，它克服了标准长短期记忆网络固有的灾难性遗忘，实现了对微尺度接触动力学的无损实时处理。

此外，我们的相位均匀采样策略有效缓解了稀疏接触事件造成的数据不平衡，大幅提升了视觉-语言-动作微调期间的样本效率。大量物理评估证实，TacMamba 完全克服了对静态帧的时间过拟合问题，而这一问题会导致纯视觉基线（例如 $\pi_{0.5}$）失效。通过严格满足 100Hz 控制回路的约束，并在避免过冲的同时保持细粒度物理精度，我们的系统能够在视觉受遮挡和长时域场景中稳健地执行任务，可靠地解决连续按压按钮、隐式状态计数和可变形衣物折叠等复杂挑战。

尽管取得了这些进展，我们的系统仍存在一定局限性。目前的配置使用一维力传感器，而非高分辨率视觉触觉传感器（例如 GelSight）。虽然这足以分离高频时间动态并避免图像处理的计算开销 [31], ，但无法提取细粒度的空间拓扑。此外，我们的物理评估并未进行全面的大规模基准测试，而是有意针对特定边缘案例（例如严重遮挡），主要与最先进的方法进行比较 $\pi_{0.5}$ [2]。尽管如此，TacMamba 的解耦特性天然支持与更广泛的开源视觉-语言-动作框架即插即用地集成。

未来的工作将通过将高维 GelSight 图像序列直接集成到 Mamba 网络中，以应对空间信息丰富的接触任务中的这些限制，并探究其促进多指灵巧手实现零样本仿真到现实迁移的潜力。

参考文献

[1] D. Kahneman，《思考，快与慢》。纽约：Farrar, Straus and Giroux，2011年。[2] K. Black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”，载于第*9*届机器人学习会议年会（*CoRL*），2025年。[3] A. Brohan等，“Rt-2：视觉-语言-动作模型将网络知识迁移到机器人控制”，载于第*7*届机器人学习会议（*CoRL*）论文集，2023年，第2165–2183页。[4] Octo Model Team等，“Octo：一种开源通用机器人策略”，载于《机器人学：科学与系统》论文集，荷兰代尔夫特，2024年。[5] G. Team等，“Gigabrain-0：一种由世界模型驱动的视觉-语言-动作模型”，*arXiv*预印本 *arXiv:2510.19430*，2025年。[6] Zhai等，“点燃视觉语言模型，迈向具身空间”，*arXiv*预印本 *arXiv:2509.11766*，2025年。[7] M. J. Kim等，“视觉-语言-动作模型微调：优化速度与成功率”，*arXiv*预印本 *arXiv:2502.19645*，2025年。[8] Z. Cheng等，“OmniVTLA：具有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。

[9] Yu等，“ForceVLA：利用力感知混合专家模型增强视觉-语言-动作模型以进行富接触操作”，arXiv预印本 arXiv:2505.22159，2025年。[10] F. Yang等，“将触觉绑定到一切：学习统一的多模态触觉表征”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2024年，第26340–26353页。[11] H. Shi等，“MemoryVLA：用于机器人操作的视觉-语言-动作模型中的感知-认知记忆”，ArXiv，第abs/2508.19236卷，2025年。[12] Li等，“MAP-VLA：面向机器人操作中视觉-语言-动作模型的记忆增强提示”，arXiv预印本 arXiv:2511.09516，2025年。[13] W. Yuan等，“GelSight：用于估计几何形状和力的高分辨率机器人触觉传感器”，《传感器》，2017年，第17卷，第12期，第2762页。[14] M. Lambeta等，“Digit：一种适用于手内操作的低成本紧凑型高分辨率触觉传感器新设计”，《IEEE机器人与自动化快报》，2020年，第5卷，第3期，第3838–3845页。[15] A. Vaswani等，“注意力就是你所需要的一切”，神经信息处理系统进展，第30卷，2017年。[16] R. Pascanu等，“论训练循环神经网络的困难”，载于国际机器学习会议。PMLR，2013年，第1310–1318页。[17] S. Bai，“通用卷积网络与循环网络用于序列建模的实证评估”，arXiv预印本 arXiv:1803 01271，2018年。[18] U. Khandelwal等，“近处清晰，远处模糊：神经语言模型如何使用上下文”，计算语言学协会（ACL），2018年。[19] X. Fang等，“KALM：利用大模型进行关键点抽象以实现物体相对模仿学习”，载于2025年IEEE机器人与自动化国际会议（ICRA）。IEEE，2025年，第8307–8314页。[20] E. Johns，“由粗到细的模仿学习：基于单次演示的机器人操作”，载于2021年IEEE机器人与自动化国际会议（ICRA）。IEEE，2021年，第4613–4619页。[21] A. Gu等，“Mamba：基于选择性状态空间的线性时间序列建模”，载于首届语言建模会议，2024年。[22] ——，“利用结构化状态空间高效建模长序列”，载于国际学习表征会议（ICLR），2022年。[23] Y. Nie等，“一个时间序列相当于64个词：基于Transformer的长期预测”，载于国际学习表征会议（ICLR），2023年。[24] T. Kim等，“面向分布偏移的精确时间序列预测的可逆实例归一化”，载于国际学习表征会议，2021年。[25] M. A. Ahamed等，“TimeMachine：一个时间序列相当于4个Mamba，用于长期预测”，载于ECAI 2024：第27届欧洲人工智能会议，第392卷，2024年，第1688页。[26] N. Wettels、V. J. Santos、R. S. Johansson和G. E. Loeb，“仿生触觉传感器阵列”，载于2008年第2届IEEE机器人与自动化学会及工程医学与生物学学会生物医学机器人与生物机电国际会议。IEEE，2008年。[27] L. U. Odhner等，“一种用于鲁棒操作的柔顺欠驱动手”，《国际机器人研究杂志》，2014年。[28] N. Fazeli、M. Oller、J. Wu、Z. Wu、J. B. Tenenbaum和A. Rodriguez，“看、感、行：基于多感官融合的复杂操作技能分层学习”，《科学·机器人学》，2019年。[29] M. J. Kim、K. Pertsch、S. Karamcheti、T. Xiao、A. Balakrishna、S. Nair、R. Rafailov、E. Foster、G. Lam、P. Sanketi等，“OpenVLA：一种开源视觉-语言-动作模型”，arXiv预印本 arXiv:2406.09246，2024年。[30] Q. Li、Y. Liang、Z. Wang、L. Luo、X. Chen、M. Liao、F. Wei、Y. Deng、S. Xu、Y. Zhang等，“CogACT：一种在机器人操作中协同认知与动作的基础视觉-语言-动作模型”，arXiv预印本 arXiv:2411.19650，2024年。[31] Q. Ye、Q. Liu、S. Wang、J. Chen、Y. Cui、K. Jin、H. Chen、X. Cai、G. Li和J. Chen，“面向类人操作灵巧性的视觉-触觉预训练与在线多任务学习”，《科学·机器人学》，第11卷，第110期，第eady2869页，2026年。