Qwen-VLA：统一跨任务、环境和机器人本体的视觉-语言-动作建模

Qwen-VLA: Unifying Vision-Language-Action Modeling across Tasks, Environments, and Robot Embodiments

Qwen Team

https://qwen.ai/blog?id=qwenvla
https://github.com/QwenLM/Qwen-VLA

摘要

机器人智能研究通常采用专门设计的模型，每个模型针对单一场景或任务，如操作或导航。这导致能力碎片化，跨任务、环境和机器人本体的泛化受限。本工作探究异构机器人决策问题能否统一在单个视觉-语言-动作（VLA）模型中处理。我们提出 Qwen-VLA，统一的机器人基础模型。用基于 DiT 的动作解码器，将 Qwen 视觉-语言建模能力从感知、理解和推理扩展到连续动作和轨迹生成。我们采用大规模联合预训练方案，跨多个数据源进行，包括机器人操作轨迹、人类第一人称演示、合成仿真数据、视觉-语言导航数据、轨迹中心监督和视觉-语言数据。为在单个共享模型中支持多机器人平台，我们引入本体感知提示条件化（embodiment-aware prompt conditioning）：机器人特定的文本描述前置于输入，指定当前的本体和控制约定。我们将操作、导航和轨迹预测统一为一个动作-轨迹预测框架。这实现跨机器人形态、任务族和环境的可迁移视觉接地、空间推理和连续动作生成。

实验在操作、导航和轨迹中心基准上进行。Qwen-VLA 支持跨任务族和机器人本体的控制，在场景布局、背景、光照、物体配置和本体变化中表现出一致的多任务性能和分布外泛化。作为统一的通用策略，Qwen-VLA-Instruct 在多个基准上达到强劲性能：LIBERO 上 97.9%、Simpler-WidowX 上 73.7%、RoboTwin-Easy/Hard 上 86.1/87.2%、R2R 上 69.0% OSR、RxR 上 59.6% SR。在真实世界 ALOHA 实验中达到 76.9% 平均分布外成功率，在 DOMINO 动态操作上达到 26.6% 零样本成功率。

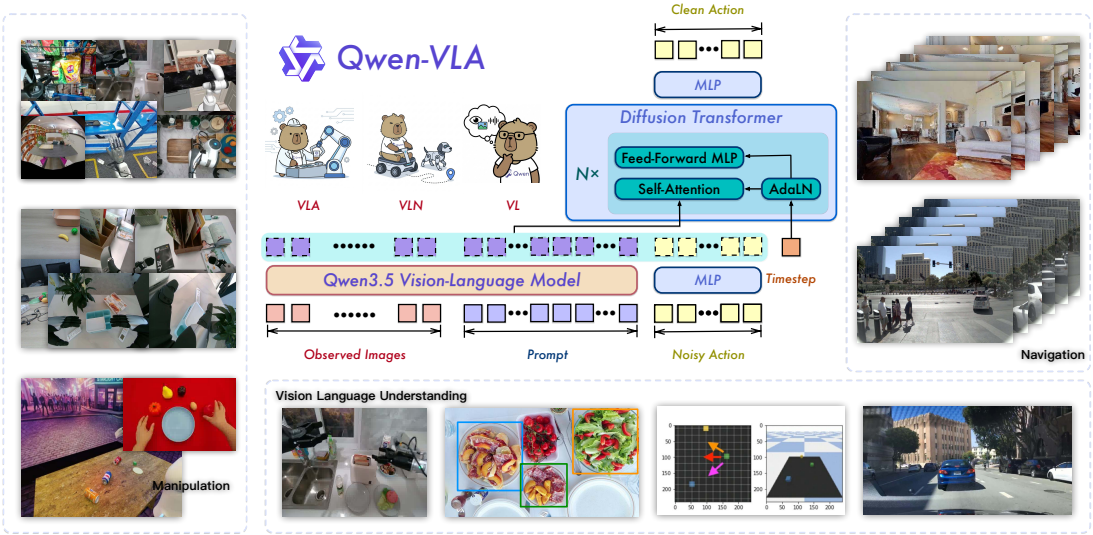

图 1: Qwen-VLA 概览，一个在混合操作、导航和视觉-语言理解数据上训练的统一机器人模型，用于生成机器人动作和文本响应。

1 引言

机器人智能 (Zitkovich et al., 2023; O'Neill et al., 2024) 的目标是构建智能体：感知物理世界、理解自然语言指令、进行空间和时间推理，并执行动作完成目标。视觉-语言模型（VLM） (Wang et al., 2024; Bai et al., 2025b; Team et al., 2025; Bai et al., 2025a; Team, 2026) 的进展大幅改进了开放世界视觉理解和语言接地。扩散型或流匹配策略 (Black et al., 2024; 2025; Intelligence et al., 2026) 在建模连续高维机器人动作上展现出强大的潜力。然而大多数现有的机器人系统 (NVIDIA et al., 2025; Luo et al., 2025; Yang et al., 2026) 仍然专门针对狭窄的任务族、机器人本体或评估设置。操作模型 (Intelligence et al., 2026) 通常针对桌面或灵巧控制训练，导航模型 (Zhang et al., 2024; 2025b; Cheng et al., 2025) 围绕室内环境中的路径点或动作预测设计。这种碎片化限制任务、环境和本体间的迁移，机器人学习难以像通用视觉-语言预训练那样扩展。

核心挑战在于本体感知决策问题表面上看起来异构。机器人操作可能需预测末端执行器姿态 (Black et al., 2024; Zheng et al., 2025; Zhang et al., 2026)、关节位置 (Fu et al., 2024; NVIDIA et al., 2025)、夹爪状态或灵巧手配置 (Luo et al., 2025)；导航可能需预测路径点或离散运动决策 (Zhang et al., 2025a; Wang et al., 2025a; Wei et al., 2025)；自我中心人类演示提供腕部和手部轨迹而非机器人控制信号。这些任务在观测格式、控制频率、预测范围、动作维度和评估协议上有差异。然而它们共享一个计算结构：本体感知智能体必须以视觉观测、语言指令和本体约束为条件，预测与任务在物理和语义上对齐的未来动作或轨迹。这一观察激发本体感知建模的统一表述。我们基于这一洞察设计联合预训练框架，将操作、导航、自我中心人类演示和视觉-语言推理纳入单个视觉-语言-动作（VLA）模型。生成的模型泛化到不同的机器人本体和任务设置。

在本工作中，我们提出 **Qwen-VLA**，统一的视觉-语言-动作（VLA）模型，用于异构本体感知决策。Qwen-VLA 基于 Qwen3.5-4B 构建，这是 Qwen 系列的本地多模态主干，在广泛的视觉-语言基准上展示最先进的性能。该主干具有强大的细粒度视觉感知、稳健的指称接地、多语言指令遵循和结构化推理能力。这些对于本体感知任务至关重要——需要解释空间关系、识别所指对象和遵循复杂多步指令。在主干之上，我们附加基于 DiT 的流匹配动作解码器，专门用于细粒度连续动作生成。Qwen-VLA 不为不同的机器人本体设计单独的输出头或特定任务架构，而是在共享的动作-轨迹预测空间中表示操作动作、导航路径点和人类自我中心运动。这使单个模型从多样化的本体感知数据集学习监督，同时保持统一的推理接口。

Qwen-VLA 的关键组成是在多样化的本体感知和视觉-语言数据上进行大规模联合预训练。预训练混合包括机器人操作轨迹、人类自我中心演示、合成仿真轨迹、视觉-语言导航数据、空间接地数据、自动驾驶视觉问答、细粒度本体感知动作说明和通用视觉-语言数据。这个混合设计覆盖低级电动机先验和高级语义推理。机器人轨迹提供直接的动作监督。自我中心人类数据提供可扩展的真实世界操作先验。导航数据引入长范围指令遵循和探索。合成仿真改进可控性和分布外覆盖。视觉-语言数据选择性采样用于本体感知相关能力（如空间推理、指称接地和指令遵循），保护并加强主干的感知和推理基础。

为实现跨本体学习，我们引入 **本体感知提示条件化**。每个训练示例都以当前机器人本体的文本描述为前缀，包括机器人平台、臂配置、控制约定、控制频率和预测范围。该提示是模型被告知本体特定控制语义的唯一接口。结合统一的动作表示，相同的动作解码器可处理不同的控制模式、动作维度和时间范围，无需改变模型架构。Qwen-VLA 在异构本体之间学习共享的视觉接地和空间推理能力，同时尊重平台特定的动作约定。

训练这样一个模型并非平凡。VLM 主干和动作解码器进入优化的不对称状态：主干已从大规模预训练中获得通用视觉-语言表示，DiT 动作解码器却随机初始化。天真地联合训练可能浪费解码器学习的计算，并破坏预训练表示。因此我们采用基于动作学习压缩观点的分阶段训练方案。语言指令如" 拿起红色杯子" 与机器人本体提示在少数几个令牌中紧凑地编码任务意图。相应的动作轨迹却跨越数百个高维关节位置值。桥接这个维度差距是结构化的解压缩问题。我们的第一个训练阶段是 **文本到动作 DiT 预训练（T2A）**，在引入任何视觉输入之前教解码器作为语言条件动作解压器。一旦这个压缩动作先验就位，后续预训练将其在视觉观测中接地。有监督微调使其特化为下游任务。强化学习优化闭环任务成功。这个方

案将动作先验压缩、视觉接地、任务特化和成功驱动的改进分离为不同的阶段。

我们在多种本体感知设置中评估 Qwen-VLA，包括机器人操作、视觉-语言导航、分布外泛化和跨本体迁移。评估既要测量领域内任务成功，也要测试统一模型是否在物体布局、光照、背景、语言指令、传感器噪声和机器人本体变化下泛化。实验表明，在异构本体感知数据上做大规模联合预训练，能获得强劲的多任务性能，超越狭隘专家训练的稳健性。这些发现支持这样的观点：操作、导航和轨迹中心本体感知任务可视为共享动作-轨迹预测问题的不同表现。

我们的贡献总结如下：

- 我们介绍 **Qwen-VLA**，一个统一的视觉-语言-动作模型，在共享的动作-轨迹空间中表述操作、导航和自我中心动作建模。基于 Qwen3.5-4B 视觉-语言主干并配备基于 DiT 的流匹配策略头，Qwen-VLA 支持跨多个机器人平台和任务族的本体感知控制。
- 我们在异构数据源上构建大规模联合预训练混合，包括来自多个机器人的操作轨迹、自我中心人类演示、合成仿真、导航和精选视觉-语言数据。我们进一步提出本体感知提示条件化，在一个模型中统一多样化的机器人平台、控制约定和预测范围，无需单独的每本体策略。
- 我们设计了一个渐进式训练方案，包括动作预训练、多模态持续预训练、有监督微调和强化学习，以构建强大的策略模型，桥接离散视觉-语言令牌和连续动作轨迹之间的差距，改进训练稳定性和下游迁移。
- 我们在操作、导航、分布外稳健性和跨本体泛化基准上评估 Qwen-VLA。结果表明，联合训练和渐进式学习在场景、物体、光照和本体变化下改进多任务性能。

2 统一机器人模型

2.1 问题表述

我们研究广泛的本体感知决策任务族，包括机器人操作、视觉-语言导航、轨迹预测和人类自我中心动作建模。我们提议在统一的本体感知模型中处理这些任务。尽管输出格式和评估协议有差异，所有任务都要求智能体将语言在视觉观测中接地、在空间和时间背景上推理，并预测未来动作或轨迹。

我们在统一的条件预测框架中表述所有任务。在时间步 t，模型接收视觉背景 o_t、语言指令 x、本体描述 e 和可选的任务标识符 z。这里，o_t 可能包含一个或多个图像帧、视频观测或历史窗口；x 指定任务指令；e 是描述当前机器人平台和控制约定的文本提示；z 在需要时识别任务族。模型被训练以预测超过预测范围 H 的目标序列 $y_{t:t+H-1}$：

$$p_\theta(y_{t:t+H-1} \mid o_t, x, e, z).$$

目标序列 $y_{t:t+H-1}$ 是任务相关的，但在统一的动作-轨迹空间中表示。对于操作任务，它对应于未来机器人动作，如末端执行器位置；对于导航任务，它表示导航决策或路径点；对于轨迹中心任务（如自动驾驶或运动预测），它表示连续坐标空间中智能体或周围实体的精确未来空间轨迹；对于自我中心本体感知数据，它捕获在结构化姿态空间（如 MANO (Romero et al., 2017) 或骨架关节序列）中表示的人体或手部运动轨迹。这种统一表述使得在单个模型内跨异构本体感知数据集进行联合优化成为可能，促进跨任务族的可迁移视觉接地、空间推理和动作生成。该表述也沿输入和输出轴可扩展：在输入侧，用外显记忆或持久状态增强条件背景 o_t 将使长范围规划和失败恢复成为可能；在输出侧，与动作一起共同预测未来视觉状态将把动作生成与世界建模统一起来，允许智能体预期其动作的后果。

2.2 模型架构

我们的模型包括一个用于高层级理解和推理的视觉-语言主干以及一个用于细粒度动作生成的流匹配动作专家。

视觉-语言主干。 我们采用 Qwen3.5 (Team, 2026) 作为主干。Qwen3.5 是本地多模态模型，通过早期视觉-语言融合训练。来自具有空间合并的 ViT 的视觉令牌直接交错到文本令牌流中，在单个 Transformer 内统一处理图像、视频和语言。其混合注意力设计在大多数层中结合门控线性注意力和定期组查询 softmax 注意力，高效编码长多模态序列，同时在需要时保留完整精度的全局推理。

动作专家。 我们附加单流 DiT 风格 (Esser et al., 2024) 的流匹配策略作为动作专家，用于预测机器人和人类本体感知数据上的精确动作 (Janner et al., 2022; Chi et al., 2023; Liang et al., 2023; Black et al., 2024)。动作专家将 VLM 隐藏状态与噪声动作块连接到一个序列中，通过具有 AdaLN 时间步条件化 (Peebles & Xie, 2023) 和与主干对齐的多部分 RoPE 的联合自注意力处理它们。这种分解设计使动作专家专门化细粒度动作生成，自然处理本体感知动作分布的多模态和高频动态，同时保留主干的预训练能力。专家用流匹配目标训练 (Lipman et al., 2023)，通过推理中的几个欧拉积分步骤产生动作序列，实现低延迟实时控制。总体而言，我们的动作专家包含约 1.15B 参数：16 个 DiT 块占大部分（每个 70.8M，共 1.13B），其余参数分布在动作投影 MLP（在原始动作维度和 DiT 隐空间之间映射，4.9M）、将 VLM 隐藏状态转换为 DiT 通道维度的线性层（3.9M）、时间步嵌入（2.8M）和输出 AdaLN 调制（4.7M）。

2.3 本体感知提示条件化

为支持共享模型中的多个机器人本体，我们在每个训练样本前添加机器人特定文本提示，说明当前平台、臂部配置和控制约定。提示遵循如下模板：

> 机器人为 {robot_tag}，配备 {单臂/双臂}[、腰部][以及移动底座]。控制频率为 {FPS} Hz。请预测执行以下任务所需的后续 {chunk_size} 个控制动作：{ori_instruction}。

机器人标签和可选修饰符（腰部、移动底座）由机器人本体设置；FPS 和 chunk_size 反映数据集的原始控制频率和预测时间范围。表 2 在 Section 3.2 中总结了预训练语料库覆盖的代表性机器人平台、臂部配置和动作类型。

2.4 统一动作和轨迹表示

我们统一了张量接口和掩码方案，但不强制所有机器人本体共享单一物理动作语义空间。每个数据集保持原生控制约定，通过机器人本体提示和数据集特定归一化来指定。每个训练样本贡献一个目标张量 $\mathbf{Y} \in \mathbb{R}^{H \times K}$，其中 H 为固定预测时间范围，K 为所有控制模式间共享的固定通道维度。

控制信号类型。 我们涵盖了两类连续控制信号。操作信号包括末端执行器相对位置（$\Delta x, \Delta y, \Delta z$）、用欧拉角或四元数表示的末端执行器旋转、绝对关节位置、夹爪开度以及灵巧手的关节角。导航轨迹信号遵循 VLN 约定，每个路径点表示为 $(\Delta x, \Delta y, \Delta\theta)$，编码地面上的相对位移和航向变化。尽管物理语义不同，这两类信号都是在时间范围内预测的实数向量序列，因此由动作专家以相同方式处理。

通道布局。 给定的控制模式使用 $c \leq K$ 个通道。这 c 个任务相关的值放置在 $\mathbf{Y}$ 的前 c 个维度中，其余 $K - c$ 个维度零填充。每通道二进制掩码 $\mathbf{M} \in \{0,1\}^{H \times K}$ 记录各通道是否携带有效信号。当且仅当 $k < c$ 且时间步 h 落在任务分块长度 $H_{\text{task}} \leq H$ 内时，$M_{h,k} = 1$。该方案无需机器人本体特定输出头；单套 DiT 参数处理所有控制模式，掩码防止填充条目影响梯度。

任务感知条件化。 每个训练样本以 Section 2.3 中描述的机器人本体感知提示作为前缀，该提示指定机器人平台、臂部配置、控制频率和预测时间范围。VLN 样本的提示还陈述导航约定和路径点时间范围。VLM 骨干网络处理这些提示 token，其隐状态与嘈杂动作分块连接作为 DiT 输入，动作专家因此始终以当前样本精确控制规格为条件，无需架构改变。

2.5 训练目标

我们端到端训练整个模型，采用两个加权目标函数：连续动作生成和视觉-语言理解。

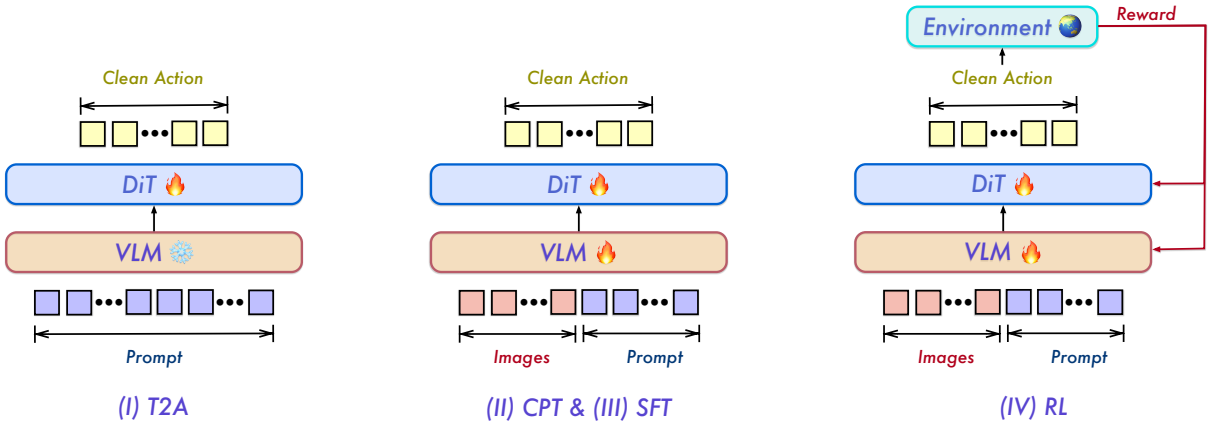

图 2: Qwen-VLA 的训练方案。阶段 I（T2A）训练 DiT 动作解码器仅从文本中重建动作，在没有视觉输入的情况下构建结构化的动作先验。阶段 II（CPT）解冻两个模块以在视觉观测中接地这个先验。阶段 III（SFT）分支为多任务和真机轨道，阶段 IV（RL）通过环境奖励优化闭环任务成功。

流匹配动作损失。 对于具有连续控制目标的样本（操作、VLN 轨迹路径点及动作对齐的人类自我中心数据），我们用条件流匹配目标 (Lipman et al., 2023) 监督动作专家。给定清晰目标 $\mathbf{Y}_0 \in \mathbb{R}^{H\times K}$ 和噪声 $\mathbf{Y}_1 \sim \mathcal{N}(0,\mathbf{I})$，我们形成线性插值 $\mathbf{Y}_\tau = (1-\tau)\mathbf{Y}_0 + \tau\mathbf{Y}_1$（$\tau \in [0,1]$）。专家 v_θ 学习预测条件速度场。

为防止填充主导梯度，我们采用*每通道、每时间步*损失，进行两级平均。对于激活 c 个通道和 H_{task} 个时间步的样本，设 $\mathbf{M} \in \{0,1\}^{H\times K}$ 为 Section 2.4 中定义的有效性掩码。先计算每个活跃通道 $k < c$ 的均方误差：

$$\ell_k = \frac{\sum_{h=1}^{H} M_{h,k} \left\| \left(v_\theta(\mathbf{Y}_\tau, \tau \mid o_{1:t}, x, e, z) - (\mathbf{Y}_1 - \mathbf{Y}_0)\right)_{h,k} \right\|_2^2}{\sum_{h=1}^{H} M_{h,k}}, \tag{1}$$

再在 c 个活跃通道上均匀平均：

$$\mathcal{L}_{\text{act}} = \mathbb{E}_{\tau,\mathbf{Y}_0,\mathbf{Y}_1}\left[\frac{1}{c}\sum_{k=0}^{c-1}\ell_k\right]. \tag{2}$$

两级平均确保每个控制维度对梯度贡献相等，与机器人本体使用通道数无关，同时完全排除填充位置。推理时通过欧拉积分从 $\tau = 1$ 至 $\tau = 0$ 生成动作分块。

视觉-语言损失。 为保留和增强骨干网络多模态能力，我们在辅助视觉-语言数据、细粒度具身动作描述、自动驾驶 VQA 和通用 VL 预训练语料库上使用标准下一个 token 预测损失：

$$\mathcal{L}_{\text{vl}} = -\sum_i \log p_\theta(w_i \mid w_{<i}, o_{1:t}), \tag{3}$$

其中 w_i 是文本 token。该目标在具身联合训练中稳定语言定位，避免感知和推理技能遗忘。

联合目标。 总体训练损失为加权组合：

$$\mathcal{L} = \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \lambda_{\text{vl}}\mathcal{L}_{\text{vl}}, \tag{4}$$

权重 $\lambda_{\text{act}}, \lambda_{\text{vl}}$ 调优以平衡两目标梯度幅度。每个小批按固定采样比例混合所有任务族样本，使每个优化步骤用操作、VLN 轨迹和视觉-语言信号联合更新骨干网络和动作专家。

3 大规模联合预训练

3.1 训练方案

可用的 VLA 模型需要联合训练认知主干和电动机解码器。这种分工类似于生物运动控制中大脑皮层和小脑的互补角色。然而在实践中，这两个模块进入训练时处于深度不对称的状态：VLM 主干已被强烈预训练，DiT 动作解码器却随机初始化。从这一点天真地联合训练是低效和不稳定的。解码器必须同时学习多个事项：动作分布的形状、如何在语言和本体上条件化、其自身参数化的流匹配动态、如何在视觉中接地动作。同时每一步都要付出编码图像的代价。来自新解码器的嘈杂梯度可能在解码器学到有用的动作结构之前扰乱预训练的主干。

我们的分阶段方案由动作学习的 *压缩* 观点驱动。原始动作轨迹是稠密的、高频的和本体相关的：单个操作事件可能包含数千个关节位置值分布在数十个自由度中。然而底层任务意图被紧凑地捕获为语言指令（" 拿起红色杯子"）和指定机器人平台的本体提示与控制约定。这个描述适合少数几个令牌。压缩的任务描述和完整动作信号之间存在巨大的维度差；桥接它是结构化的 *解压缩* 问题。

我们将 T2A 视为学习这个解压缩映射。通过扣留图像并仅在语言条件动作预测上训练 DiT，我们强制解码器获得动作空间上的结构化先验，该先验完全由语言索引。这不仅仅是热启动：解码器学习不同的语言描述如何选择动作分布的不同区域、本体提示如何将相同的任务意图调制为平台特定的电动机程序、完整动作轨迹在序列级别上的时间一致性和可组合性。有了这个语言索引的动作先验就位，后续的多模态训练可以将其容量集中在具体视觉观测中接地先验，而不是从头学习动作生成。

基于这一原则，我们在预训练的 Qwen3.5 VLM 主干上采用四阶段训练方案：(I) **文本到动作 DiT 预训练（T2A）**；(II) **持续预训练（CPT）**；(III) **有监督微调（SFT）**在两个并行分支中；(IV) **强化学习（RL）**。每个阶段由它在前一阶段中关闭的差距定义。

阶段 I：文本到动作 DiT 预训练（T2A）。 我们冻结 VLM，仅训练 DiT，以文本和本体提示 e (Section 2.3) 为条件，有意不提供图像。解码器作为纯语言到动作解压器运行，从紧凑的语言编码单独重建高维动作分布，没有任何视觉快捷方式。T2A 安装结构化的动作先验：语言选择动作空间的区域，本体提示指定平台特定的电动机参数化，流匹配动力学在引入任何图像之前管理生成过程。

阶段 II：继续预训练（CPT）。 在解码器热启动的情况下，CPT 将其多模态容量集中在 T2A 无法解决的问题上：在视觉观测中接地动作并使主干适应本体感知感知。我们解冻两个模块并在异构混合上训练 (Section 3.2)，故意结合仿真和真机轨迹，使生成的检查点已见两个域。这个特性对于后续后训练阶段很重要：SFT 和 RL 可专门针对任一域。

阶段 III：有监督微调（SFT）。 虽然 CPT 提供了广泛的跨本体和跨任务覆盖，下游部署受益于将模型与精选的、从目标任务分布中提取的高质量演示对齐。我们从 CPT 检查点将 SFT 分为两个并行轨迹。第一条轨迹执行多任务 SFT，在异构任务上联合微调模型，包括视觉问题回答、空间接地、操作和导航，在本体平衡和任务平衡采样下。第二条轨迹从 CPT 检查点在遥操作数据上微调，用于真实世界机器人部署，测试 CPT 的跨域先验是否转移到物理硬件。

阶段 IV：强化学习（RL）。 SFT 在演示上优化似然性；最终重要的是闭环任务成功，这是执行轨迹的一个特性，任何模仿目标都无法直接优化。从多任务 SFT 检查点开始，RL 用收集的稀疏二元成功奖励微调策略，专门在单个仿真环境（SimplerEnv）中，产生最终模型 Qwen-VLA-Instruct。这个故意的狭隘 RL 设置测试任务成功改进在一个仿真环境中获得是否可转移到其他分布外环境。

3.2 预训练数据

预训练语料库的质量和多样性直接决定认知骨干网络和动作解码器在不同机器人本体和任务族间的共适应程度。我们构建大规模异构预训练混合数据集，涵盖具身感知、空间推理和动作生成的广泛能力。混合数据集跨越五个数据族：机器人操作轨迹、人类自我中心演示、合成仿真数据、导航及轨迹为中心的

数据，以及辅助视觉-语言数据。表 1 总结各数据源的组成和采样权重。

表 1: 预训练数据混合成分。

数据源	比例（%）
机器人操作轨迹	74.2
人类自我中心轨迹	6.0
导航轨迹	7.5
合成仿真轨迹（本文）	3.7
通用视觉-语言数据	3.4
空间定位（2D）	2.5
自动驾驶 VQA	2.4
细粒度具身动作描述	0.2
总计	**100.0**

3.2.1 机器人操作轨迹

真实和仿真机器人操作轨迹构成预训练语料库的核心，占约 74.2%。数据涵盖桌面操作、移动操作、双臂任务和各种本体的灵巧手控制。

公开数据集。 为了改进数据规模、多样性和跨本体泛化，我们在公开真机数据集的广泛混合物上训练，包括 RobotSet (Bharadhwaj et al., 2023)、Galaxea (Jiang et al., 2025)、AgiBot World (AgiBot-World-Contributors et al., 2025)、RoboCOIN (Wu et al., 2025b)、RoboMIND V1/V2 (Wu et al., 2025a; Hou et al., 2025)、RDT-1B (Liu et al., 2025a)、DROID (Khazatsky et al., 2024)、BridgeData V2 (Walke et al., 2024)、RH20T (Fang et al., 2023)、RT-1 (Brohan et al., 2022) 和 BC-Z (Jang et al., 2022)。这些数据集覆盖桌面操作、移动操作、双臂操作、灵巧手控制和野外任务，总超 10000 小时跨异质本体和场景的交互数据。大规模混合物在机器人形态、视点、物体、背景、语言指令和动作分布上提供互补监督，减少单一本体或环境的过拟合，增强分布偏移下的鲁棒性。

除真机数据外，还包含 InternData-A1 (Tian et al., 2025) 和 GR00T-X-Embodiment-Sim (NVIDIA et al., 2025) 的仿真轨迹，通过多样虚拟环境中的运动规划生成。这些轨迹增强场景和物体多样性，特别是长尾配置和任务布局。训练前，所有真实和仿真数据集整理并转为统一的观测-动作格式，保留原始任务语义，实现跨异质机器人经验的可扩展训练。

专有数据集。 补充了超 1000 小时内部采集的真机轨迹，涵盖各种操作任务和机器人平台，约占预训练混合物的 20%。此外，用 Section 3.2.3 中说明的可扩展仿真流程生成超 800 万条综合轨迹，占混合物的 3.7%，大幅扩展场景、物体和任务配置多样性，超过纯真机遥操作范围。

本体感知的提示条件化。 如 Section 2.3 所述，每个训练示例前有机器人特定提示，指定当前平台、臂配置和控制约定。表 2 列出预训练语料库中代表性本体及其臂配置和动作类型。

动作表示。 不同数据集采用不同动作约定：有的提供笛卡尔空间绝对末端执行器位姿 (Xie et al., 2026)，有的提供增量末端执行器命令，有的用绝对或相对关节空间控制。保留每个数据集的原始动作格式而非转换为共享表示，通过本体感知提示条件化指示模型当前控制约定。

各动作维度按数据集用分位数统计归一化。对数据集 k 的各动作维度 d，计算第 1 和 99 分位数 q_{01}^k 和 q_{99}^k，应用线性映射：

$$\tilde{a}_d = 2 \cdot \frac{a_d - q_{01}^k}{q_{99}^k - q_{01}^k} - 1, \tag{5}$$

将结果裁剪到 $[-1, 1]$。按数据集的分位数归一化消除跨本体和动作空间的规模差异，同时保留每个源内的相对运动结构。

表 2: 预训练语料库中的代表性机器人本体。“EEF” = 末端执行器位姿；“Joint” = 关节角度；“Δ” = 增量（相对）命令；“Abs” = 绝对命令；“G” = 夹爪；“DH” = 灵巧手。

机器人	臂数量	动作类型
WidowX	Single	ΔEEF + G
Google Robot	Single	ΔEEF + G
Franka Panda	Single / Dual	ΔEEF + G; Abs Joint + G
ARX5	Dual	ΔEEF + G
Fourier GR-1	Dual	ΔEEF + G
Mobile ALOHA	Dual	ΔEEF + G; Abs Joint + G
AgiBot A2-D	Dual	Abs Joint + G; Abs Joint + DH
Galaxea R1	Dual	Abs Joint + G
AIRBOT MMK2	Dual	Abs Joint + DH
TienKung	Dual	Abs Joint + G; Abs Joint + DH
Real Human	Dual	ΔEEF (from MANO)

语言指令。 语言指令来自原始数据集注释和模型生成字幕的组合。所有指令经过质量过滤和一致性检查流程，语言注释与观测运动不一致的轨迹被丢弃，确保跨多样源的可靠语言-动作对齐。除动作标记数据外，还结合细粒度具身视频-字幕对（如 Section 3.2.5 所述）作为辅助视觉-语言监督。密集语言信号填补粗粒度任务标签与丰富机器人行为语义描述的差距，使模型能在推理时泛化到更广泛的自然语言指令。

相机视角表示。 许多机器人数据集从多个相机视点提供观测，如头部自中心相机和一个或两个腕部相机。为向模型显式表示视角来源，在令牌流中用一对视角特定的边界令牌包装各图像：

$$\texttt{<|}\textit{tag}\texttt{\_start|>}\ \langle\text{image}\rangle\ \texttt{<|}\textit{tag}\texttt{\_end|>}$$

其中 *tag* 标识相机源，如 `ego`、`cam_left_wrist` 或 `cam_right_wrist`。显式视角标签让视觉-语言模型骨干形成视角感知表示，动作专家在生成动作时有选择地关注各视点，无需架构改变或额外输入通道。

数据清洁。 应用多阶段过滤流程移除损坏或缺失帧的轨迹、接近零方差的动作序列（静态录制）和长度异常的片段。对无显式动作标签的数据集，通过本体感知状态序列的有限差分恢复伪动作。

3.2.2 自我中心人类数据

自我中心人类演示比遥操作机器人轨迹提供更丰富且可扩展的真实世界操作经验。人类每天与开放世界环境中的各种物体相互作用，自然产生跨越比机器人遥操作更广泛的场景、物体和任务语义的灵巧操作行为。最近的研究表明，在大规模自我中心人类视频上训练能赋予视觉-语言-动作模型更丰富的操作先验，并改进对下游机器人任务的泛化 (Kareer et al., 2025; Luo et al., 2025; Li et al., 2026b; Luo et al., 2026; Zheng et al., 2026; Hu et al., 2026)。受这些发现的激励，我们整合多样化的自我中心人类操作数据集（占总预训练混合的 6.0%），为模型提供补充机器人轨迹数据的广泛操作先验。

数据来源。 我们的自我中心语料库来自四个公开可用的数据集，每个数据集在任务多样性、场景变异性和注释粒度方面都提供补充性覆盖。**(1)** 我们采用由 VITRA (Li et al., 2026b) 处理的 Ego4D (Grauman et al., 2022) 和 EPIC-KITCHENS (Damen et al., 2022) 子集，该工具应用自动化管道将自我中心人类视频分割成原子操作轨迹并生成细粒度语言注释，以及逐帧 3D 手部和摄像机运动轨迹。**(2)** EgoDex (Hoque et al., 2026) 是用 Apple Vision Pro 捕获的大规模灵巧操作数据集，包含 829 小时自我中心视频，在 194 个不同的桌面任务中配对 3D 手部和手指跟踪。**(3)** EgoVerse (Punamiya et al., 2026) 是一个协作自我中心演示平台，跨越 1,300 小时超过 1,965 个任务和 240 个场景，提供标准化格式和操作相关注释。**(4)** Xperience (Ropedia, 2026) 是大规模自我中心多模态数据集，提供同步第一人称录制、深度、手部和身体运动捕捉以及分层语言注释。

动作表示。 模型在任意时间步接收自我中心图像以及语言指令，并预测包含在固定预测范围内的未来双臂腕部运动和手部铰接的动作块。对于每只手，腕部运动表示为相对于初始帧的未来时间步的腕部坐标

系的 SE(3) 变换。在训练期间，这个相对动作用平移向量和轴角旋转表示，每只手产生 6 个腕部动作维度。为紧凑表示手部铰接，我们在所有人类数据集的手部 45 维轴角关节姿态上执行主成分分析（PCA），保留前 10 个主成分的权重。这些低维系数称为本征抓握 (Ciocarlie et al., 2007; Yuan et al., 2025)，捕获人类手部姿态变异的主导模式，同时丢弃关节冗余。模型预测相对腕部动作和每只手 10 个本征抓握系数，对于自我中心人类数据，每个时间步产生总共 32 个动作维度。

3.2.3 合成仿真数据

我们构造大规模合成仿真数据管道来改进机器人感知监督的覆盖、可控性和稳健性。管道包含两个补充性组件：(1) *视觉-语言-动作数据*，模型从任务指令和图像观测中预测动作；(2) *语言-动作数据*，模型仅从语言中预测动作。这两个监督源起不同但协同的作用。仅文本组件鼓励模型获得高级任务抽象和语言-动作规律性，无需依赖视觉外观。视觉条件化组件在真实感知观测、场景变异和机器人化交互动力学中接地这些抽象。我们基于 IsaacLab (Mittal et al., 2025) 开发仿真环境管道，基于 cuRobo (Sundaralingam et al., 2023) 用于碰撞避免运动规划。

视觉-语言-动作数据。 对于视觉条件化合成数据，我们使用 RoboInF (XLANG Lab, 2026) 的内部早期版本生成标准 VLA 风格监督，模型接收语言指令和图像观测作为输入，预测机器人动作作为输出。RoboInF 构造操作场景、生成场景条件化任务、合成可执行的成功检查、通过模拟器反馈产生机器人运动程序、在域随机化下展开成功轨迹。完整技术细节在我们的 RoboInF 博客文章中提供，这里只总结数据集组成和多样性。

为简便起见，我们专注于随机放置场景生成设置。我们构造 20 个桌面场景，每个场景配置有 10 种不同的物体初始姿态配置，产生 200 个基础场景配置。这些场景包含排列在物理有效桌面布局中的多样化日常物体，为操作提供多样的视觉和空间背景。

在这些场景之上，我们生成 450 个操作任务，跨越短期和长期行为。短期任务通常涉及少量原始操作步骤，例如 *“并排放置两个绿色订书机”*。长期任务需要多个顺序交互或更多组合推理，例如 *“把饮料放在一起，把清洁海绵单独放在一边”*，其中机器人需要按顺序操作多个物体。如图 3 所示，这个任务混合使模型同时接触原子操作技能和扩展时间范围上的更高级指令遵循模式。

对于每个任务，我们生成 300 条成功轨迹，具有不同的环境和执行增强。在展开期间，我们随机化视觉、几何和控制相关因素：照明、摄像机姿态、背景、桌纹理、机器人初始状态、物体初始姿态和控制器动力学（如刚度和阻尼）。对于视觉随机化，我们从约 3K 个背景和 1K 个桌纹理的大型池中采样。对于其他因素如摄像机姿态、机器人状态、照明和控制器参数，我们手动定义合理范围并采样这些范围内的变异。这些增强用于提升模型对外观、视点、布局和低级执行动力学分布偏移的稳健性。

这些生成数据的有用特性是轨迹从运动规划程序产生，这些程序自然分解为中间阶段。我们另外将每个完整轨迹分割成多个子任务轨迹，如图 3 所示。这在多个时间粒度上提供监督：完整轨迹教模型在较长范围上遵循高级指令，子任务轨迹使模型接触较短的原子行为和中间目标。这个混合帮助模型更好地连接高级语言指令与完成它们所需的较低级动作段。

总体而言，视觉条件化合成数据包含 359,848 条完整成功轨迹（包括子任务段）。生成的数据集提供大规模多样化、可控和自动验证的 VLA 监督来源，补充真实世界演示并改进模型在任务类型、视觉条件和执行变异上的泛化。

语言-动作数据。 我们另外构造纯文本动作数据集作为补充监督来源。此组件的目标是在引入视觉观测之前将模型暴露于广泛的操作意图和动作模式空间，从而鼓励语言-动作联合空间中的泛化。我们设计可由模拟器中机器人执行的多样化原子和组合操作任务，并使用无图像输入的机器人状态中心管道生成对应轨迹。此数据主要作为语义和行为预训练形式，帮助模型仅从语言指令捕获常见任务结构：移动、分组、分离、对齐、旋转、堆叠和重新排序物体。

我们定义六个任务模板系列，跨越核心单臂操作原始体：拾取-放置、线性推动、线性拉拔、带重新定位的旋转、朝向视点方向的旋转和两个物体的位置交换。每个模板在随机化空间参数和自然语言描述

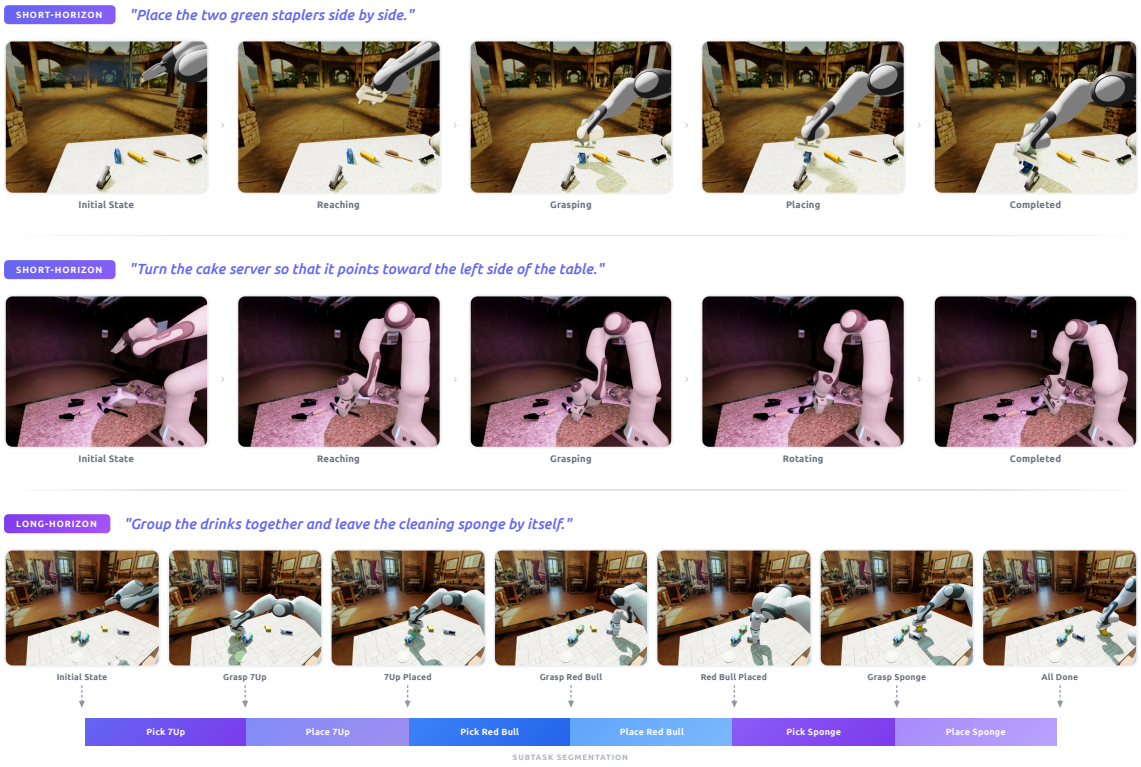

图 3: 通过 RoboInF 生成的数据示例。顶行显示短期任务“并排放置两个绿色订书机”，包含触及、抓取、运输和放置的紧凑序列。底行显示长期任务“把饮料放在一起，把清洁海绵单独放在一边”，需要多个物体操作，可分解为拾取和放置每个饮料等子任务段。

的同时编码规范运动结构。指令通过从模板特定的动词、前置词、物体名称和空间参考词汇表中程序性组合来产生相同基础行为的多样化描述（例如“把苹果放进容器”、“把海绵移向目标位置”）。此设计在保持足够简单以扩展到数百万轨迹的同时，提供对任务覆盖和难度的精确控制。

为确保对本体特定运动学和工作空间几何体的广泛覆盖，我们在六个单臂机器人配置（Franka Panda、UR10e、UR5e、Kinova Gen3、TM12 和 xArm7）上实例化所有六个任务模板，跨越不同的自由度、臂形态和可达工作空间。对于每个机器人-任务对，我们生成约 200k 条轨迹，总共产生约 7.2M 条轨迹和超过 14,000 小时的模拟机器人轨迹数据。轨迹合成在没有物理模拟或场景渲染的情况下进行：对于每个实例，我们在机器人可达工作空间范围内采样末端执行器目标姿态，随机化位置、方向和姿态间距离，并使用 GPU 加速批量运动规划器计算无碰撞关节空间路径。生成管道在多个 GPU 上并行化，每个工作进程独立采样目标配置、规划运动和写入成功轨迹。仅保留运动学有效的轨迹，其中规划器在所有需要的运动段中产生可行、无碰撞的解决方案。

每个轨迹以 50 Hz 记录关节位置、关节速度、末端执行器姿态（位置和方向）和夹爪状态，提供可支持训练期间各种动作表示和控制模式的丰富灵活监督信号。

在训练期间，此数据集作为第 I 阶段 T2A 预训练的主要语料库（Section 3.1），其中 DiT 动作解码器在引入视觉接地之前初始化为语言-动作对应关系。通过在没有视觉条件的情况下向解码器暴露大量多样化语言-动作对，此组件预先排除视觉快捷方式并建立强大的动作先验，该先验转移到下游视觉条件化训练阶段并补充之。

3.2.4 导航数据

我们整合导航数据（7.5

指令遵循。 指令遵循导航数据（4.3

物体搜索。 物体搜索数据（2.3

目标跟踪。 目标跟踪数据（1.0

3.2.5 Vision-language data.

进一步加入辅助视觉-语言监督（8.5%）以强化语义接地、细粒度指令跟随和通用视觉推理。

细粒度具身动作字幕。 大多数机器人数据集仅提供粗粒度任务标签（如" 拿起陶瓷碗"），不足以进行精确动作预测：相同标签可对应完全不同的执行策略（如从左 vs 右、顺时针 vs 逆时针），在策略学习中造成歧义。受视频生成中的细粒度文本控制启发，构建密集动作描述，将语言与各片段紧密对应。为建立多样但可管理的注释集，从主要开源操作数据集各任务类别中抽取片段，应用基于聚类的选择策略：按动作信息分组，采样代表性片段确保广泛覆盖。各选定片段沿 13 个维度注释：动作原语、演员身份、物体识别和消歧、接触区域、源和目标位置、轨迹和方向、夹爪状态和身体运动。注释采用两阶段流程。第一阶段，视觉-语言模型（Qwen3.6-plus）观看视频并提取粗动作序列和被操纵的物体。第二阶段，将视频密集采样成帧以保留细微运动细节，模型在第一阶段输出条件下重新观看这些帧并生成细粒度逐步描述，充实接触点、空间关系和运动轨迹。对多视角数据集，额外精化过程结合腕部或侧摄像头录像以纠正仅从近距离或替代视角可见的细节。所有生成字幕由人类注释者最终审查和纠正。流程产生约 48000 条细粒度视频-字幕对（占预训练混合物的 0.2%）。表 3 对比粗标签与为相同片段提供的细粒度字幕。

表 3: 相同片段的粗标签对比细粒度动作字幕。

粗标签	" 拿起、旋转并放置陶瓷碗。"
细粒度	*步骤 1*：从右侧远端拿起陶瓷碗。 *步骤 2*：顺时针旋转碗两整圈。 *步骤 3*：将碗放在桌子中心。

基于注释数据，进一步微调专用视觉-语言模型用于具身动作字幕。同时设计配套评估基准，包括从真值注释派生的问答对和字幕质量评分标准。

自动驾驶视觉问答。 结合自动驾驶视觉问答数据（2.4%）作为视觉接地和决策导向监督的补充源。此混合物针对四种可转移能力：（1）*时间场景理解*——驾驶场景在自我运动、动态代理和变化危险下演进。纳入 LingoQA (Marcu et al., 2024)、DriveAction (Hao et al., 2025) 和 MMAU (Fang et al., 2024) 以强化事件理解和未来状态预测。（2）*环视空间推理*——多摄像头观测提供宽视野几何背景。纳入 Impromptu-VLA (Chi et al., 2025)、nuScenes-QA (Qian et al., 2024)、nuScenes-MQA (Inoue et al., 2024)、MapLM (Cao et al., 2024) 和 WaymoQA (Yu et al., 2025b) 以提升跨多样车辆、传感器和道路布局的视角鲁棒定位和空间关系理解。（3）*语言接地定位*——具身代理需在行动前将语言与物体、区域和空间关系对应。纳入 CODA-LM (Chen et al., 2025a)、Talk2Car (Deruyttere et al., 2019)、DrivingVQA Corbiere et al. (2025)、DriveLM (Sima et al., 2024)、W3DA (Zhou et al., 2025) 和 GRAID (Elmaaroufi et al., 2025) 实现物体级接地和区域感知问答。（4）*规划感知推理*——安全决策需将感知与路线、车道和轨迹约束关联。纳入 Bench2Drive-VL (Jia et al., 2026)、DriveGPT4 (Xu et al., 2024)、OmniDrive (Wang et al., 2025b)、Senna (Jiang et al., 2024)、NAVSIM-RecogDrive (Li et al., 2026c) 和 NAVSIM-Traj 监督机动选择和未来轨迹预测。预训练前，所有数据源标准化为统一对话格式。多选注释转为自由形式问答监督，边界框规范化为 1000 尺度坐标系。多帧样本加帧标签以维持时间顺序，环视样本分配视角标签以显式表示摄像头身份。这些结构化源为物体中心接地、时间状态估计、意图预测和目标条件决策提供可转移监督，充分补充了操作和导航数据。

空间接地。 包含 2D 边界框接地数据（2.5%）以强化物体级空间理解。准确的空间接地是语言条件操作的基础，模型需从指令指定的描述中定位任务相关物体。

通用视觉-语言数据。 为在具身视觉-语言-动作模型预训练中保持鲁棒视觉感知和语言接地，用精心策划的通用视觉-语言数据混合物（总 token 的 3.4%）补充主要的动作中心数据集。此混合物包括：（1）图像-文本对齐的字幕数据；（2）世界知识保留的知识导向视觉问答和推理；（3）细粒度文本感知的 OCR 和文本接地样本；（4）跨模态对齐的交错指令跟随数据；（5）物体级语言接地的常见任务（如指称表达、空间关系预测）。为更好支持具身推理，策略性地在此混合物中加权视频中心、空间关系和 3D 感知监督。这些通用视觉-语言目标缓解灾难性遗忘、保留基础世界知识先验，增强语言条件具身感知的鲁棒性。

4 后训练

T2A 和 CPT 的大规模预训练产生 **Qwen-VLA-Base**，一个表现出广泛跨任务和跨本体泛化的通用视觉-语言-动作模型。虽然这种广泛覆盖为模型配备了多才多艺的知识，但它尚未产生在特定下游任务上进行可靠闭环控制所需的精度。为了弥合这个差距，我们介绍了一个两阶段后训练程序，将基础模型专门化用于精确任务执行：

(i) 多任务有监督微调（SFT）阶段，在异构任务上联合微调视觉-语言主干和动作专家，包括视觉问答、空间接地、操作和导航，在本体平衡和任务平衡采样下。

(ii) 强化学习（RL）阶段，从 SFT 检查点初始化，通过直接针对从仿真中的在策略展开获得的任务成功驱动奖励进行优化来进一步完善策略，产生最终模型 **Qwen-VLA-Instruct**。

在两个阶段中，我们使用余弦衰减学习率，视觉-语言主干和动作解码器有单独的分组计划，以及与预训练一致的梯度裁剪。每个阶段的具体数据和方法在后续章节中详细说明。

4.1 多任务有监督微调

数据。 SFT 数据混合包括三个类别：通用视觉-语言样本、视觉-语言导航插曲和机器人操作演示。首先，我们精选一组视觉-语言样本——涵盖视觉问答、空间接地和动作标题——在微调过程中保持主干的视觉理解。其次，我们从仿真中收集机器人操作演示，包括跨越单臂、双臂和人形机器人的多样化平台的轨迹，所有都使用每个块 16 个动作步的预测范围 (Community, 2026)。遵循预训练中使用的相同本体感知提示设计（Section 2.3），每个训练样本都以指定机器人平台、臂配置、控制频率和预测范围的文本提示为前缀。第三，我们从连续控制环境中收集视觉-语言导航插曲，涵盖多样化的室内场景和指令风格。我们仅保留成功完成的插曲，每个块预测 8 个路径点。

目标。 SFT 阶段联合优化两个损失：视觉-语言令牌上的下一个令牌预测和动作令牌上的流匹配。损失权重设为视觉-语言下一个令牌预测的 0.1 和操作和导航动作预测的 1.0，确保模型保留其语言和视觉理解，同时将梯度容量集中在动作生成上。

4.2 强化学习

基于似然的 SFT 优化衡量策略模仿演示效果的代理。然而在机器人化控制中最终重要的是闭环任务成功：执行的轨迹是否达成目标。这两个目标可以分化：策略可能分配高似然性于合理动作，但当其自有输出将状态分布从演示支持中转移时失败。RL 阶段通过直接针对策略本身展开的轨迹上的任务成功驱动奖励进行优化来弥合这个差距，产生最终模型。我们使用 RLinf 框架实现此阶段 (Yu et al., 2025a)。

策略优化目标。 我们采用近端策略优化（PPO） (Schulman et al., 2017) 与广义优势估计（GAE） (Schulman et al., 2015) 作为 RL 算法。令 π_θ 表示当前策略，$\pi_{\theta_{\text{old}}}$ 表示生成展开批的策略。裁剪代理目标为

$$\mathcal{L}^{\text{actor}}(\theta) = -\,\mathbb{E}_t\Big[\min\Big(r_t(\theta)\,\hat{A}_t,\ \text{clip}\big(r_t(\theta),\,1-\epsilon,\,1+\epsilon\big)\,\hat{A}_t\Big)\Big], \tag{6}$$

其中 $r_t(\theta) = \pi_\theta(a_t \mid s_t) \,/\, \pi_{\theta_{\text{old}}}(a_t \mid s_t)$ 是重要性比率，$s_t = (o_t, x, e)$ 是当前状态（视觉观测、任务指令和本体感知提示遵循 Section 2.3），a_t 是范围 H 的预测动作块，$\hat{A}_t$ 是 GAE 优势估计，折扣率 γ=0.99，迹

衰减 λ=0.95，ϵ=0.2 是对称应用的裁剪阈值。总损失将策略代理与值函数回归项结合：

$$\mathcal{L}(\theta) = \mathcal{L}^{\text{actor}}(\theta) + c_v \, \mathcal{L}^{\text{value}}(\theta), \tag{7}$$

其中 $\mathcal{L}^{\text{value}}$ 是值预测的损失，c_v=1。我们为每个展开批执行四个优化轮次。

值估计。 而不是训练单独的评论家网络，我们直接将轻量级值头附加到视觉-语言主干。值头对所有 VLM 隐藏状态进行平均池化，并通过线性投影将其映射到标量值估计。我们在 VLM 隐藏状态进入值头之前应用停止梯度，以便值函数梯度不会通过预训练主干反向传播。值头用裁剪 MSE 损失和单独的学习率（10^{-4}，约为演员学习率 5×10^{-6} 的 20×）进行训练，允许其快速收敛，同时策略更新保持保守。

流匹配下的对数概率估计。 将 PPO 应用于我们的流匹配动作解码器的关键挑战是计算重要性比率 $r_t(\theta)$ 所需的对数概率 $\log \pi_\theta(a_t \mid s_t)$。与自回归令牌基策略（其中对数概率可直接从 softmax 输出获得）不同，流匹配模型通过学习的速度场和迭代去噪定义隐式密度。我们通过在每个欧拉去噪步骤注入受控噪声，将确定性概率流 ODE 转换为对应的 SDE (Song et al., 2021)，使得每个转移变成其对数概率可以在没有数值 ODE 积分的情况下分析计算的显式高斯。在展开期间我们存储中间去噪状态；在 PPO 更新处我们重新评估当前参数下的速度场并重新计算高斯对数概率，以可忽略的额外成本产生重要性比率。默认地我们为对数概率估计为每个展开随机选择单个去噪步骤，在重新计算期间仅需要一个额外的 DiT 前向传递。对数概率和优势都在*动作块*级别计算：每个 H=16 个动作步的块接收单个标量奖励和单个优势估计，匹配流匹配解码器输出的时间粒度。

奖励设计。 我们使用来自模拟器的稀疏二元奖励：如果在插曲末尾实现任务目标则 R=1，否则 R=0。跨动作块的信用分配由 GAE（eq. (6)）处理，该 GAE 将插曲级信号通过值基线反向传播。不使用学习的奖励模型；所有信号来自模拟器的地真任务完成语义。

展开基础设施。 在策略展开在镜像下游评估基准的仿真环境中收集。我们采用分离的客户端-服务器架构，其中远程基准服务器托管仿真环境，训练过程通过网络接口与其通信。这种分离允许仿真工作负载独立于 GPU 密集的策略优化进行扩展。我们实例化 N=128 个平行环境实例，分布在多个操作任务套件上，非统一分配与任务难度成比例。每个训练迭代收集 8 个展开轮次，每个 128 个环境步骤，产生 $N \times \lfloor 128/H \rfloor \times 8 = 128 \times 8 \times 8 = 8{,}192$ 转移块每次迭代（其中 H=16 是动作块长度）。在展开期间，从具有温度 τ=1.0 的策略采样动作；在评估时，我们将温度降低到 τ=0.6 以锐化动作分布。在展开期间使用的本体感知提示与 SFT 中的相同，确保 RL 阶段不重新引入提示条件化中的分布偏移。

分布外泛化。 因为本体感知提示是唯一的平台特定接口，RL 精细化策略转移到分布外环境、任务和本体，无需额外的适应头或特定域微调。在部署时，我们简单地将本体感知提示替换为物理机器人平台的文本描述，包括其臂配置、控制频率和动作空间，同时保持 VLM 主干和 DiT 动作解码器不变。这种零样本泛化由我们训练方案的两个特性启用：(1) CPT 阶段已经将模型暴露于仿真和真实世界视觉分布（Section 3.1），所以主干的视觉表示不特定于仿真渲染器；(2) RL 阶段优化对视觉域不变的任务成功指标，鼓励策略依赖语义有意义的视觉特征而不是渲染器特定纹理。我们在 Section 5.2.3中量化每个后训练阶段的累积效果。

5 实验

我们进行广泛实验以评估 Qwen-VLA 在仿真和真实世界设置中的性能，涵盖两个核心机器人化 AI 领域：机器人操作和视觉导航。我们始终评估两个模型变体：**Qwen-VLA-Base**，用大规模多样化机器人化数据预训练，和 **Qwen-VLA-Instruct**，进一步用从多样化仿真环境和任务中收集的指令遵循数据微调，能够在一个统一模型中以更大精度和可靠性执行细粒度操作和导航。

5.1 主要结果

5.1.1 仿真中的操作结果

我们在四个仿真环境上进行基准测试，全面评估 Qwen-VLA 的操作能力，涵盖单臂和双臂设置：LIBERO (Liu et al., 2023)、Simpler (Li et al., 2024)、RoboCasa-GR1 桌面任务 (Nasiriany et al., 2024) 和 RoboTwin 2.0 (Mu et al., 2024)。

- **LIBERO** 是包含四个分割的多样化任务的单臂桌面基准：Libero-Spatial、Libero-Object、Libero-Goal 和 Libero-Long。
- **Simpler-WidowX** 是使用 WidowX 机器人臂的真实到仿真评估套件。
- **RoboCasa-GR1** 是具有 24 个原子厨房任务的双臂人形基准。
- **RoboTwin 2.0** 是具有 50 个双臂任务的双臂基准，分为简单和困难难度等级。

这些基准测试覆盖从桌面拾取-放置到长期双臂厨房任务的广泛操作技能。对于所有基准，我们设置动作块长度 $H = 16$，根据 StarVLA (Community, 2026) 中的标准评估协议报告平均成功率。

表 4: **跨基准的机器人操作结果：专家 vs. 单个通用模型**。专家模型在每个基准上分别微调，而我们的通用模型（Qwen-VLA）在所有本体上联合训练一次，在所有平台上评估，无需按基准适应。**粗体**：最佳；下划线：第二最佳。

Method	Type	Benchmark				
		LIBERO	RoboCasa-GR1	Simpler-WidowX	RoboTwin-Easy	RoboTwin-Hard
π_0 (Black et al., 2024)	Specialist	94.4	–	–	65.9	58.4
StarVLA-OFT (Community, 2026)		96.6	48.8	64.6	50.4	–
GR00T N1.6 (NVIDIA et al., 2025)		97.2	49.9	63.2	47.6	–
$\pi_{0.5}$ (Black et al., 2025)		97.6	37.0	46.9	82.7	76.8
ABot-M0 (Yang et al., 2026)		**98.6**	**58.3**	–	86.0	85.0
Being-H0.5 (Luo et al., 2026)		97.6	53.3	–	–	–
Qwen-VLA-Base	Generalist	90.8	40.4	64.3	64.3	66.4
Qwen-VLA-Instruct		97.9	56.7	**73.7**	**86.1**	**87.2**

表 4 比较我们的通用模型与最先进的专家策略。专家模型在每个基准上分别微调并保持本体特定，我们的单个通用模型在跨越多个场景和本体的多样化大规模数据上联合训练，通过本体感知提示在所有四个平台上启用部署。

单个通用模型优于大多数专家。 尽管是一体式模型，Qwen-VLA-Instruct 超越大多数专家基线。在 LIBERO 上达到 97.9%，与最好的专家相当。在 RoboCasa-GR1 上达到 56.7%，超越 $\pi_{0.5}$（37.0%）、GR00T N1.6（49.9%）和 Being-H0.5（53.3%）。在 Simpler-WidowX 上达到 73.7%，超越 StarVLA-OFT（64.6%）等专家基线。在 RoboTwin-Easy/Hard 上达到 86.1%/87.2%，**超越**先前最好的专家 ABot-M0（86.0%/85.0%）。这些结果确认联合多本体训练不牺牲任务特定性能；通用模型甚至超越专用模型。

预训练提供强大基础，指令调优产生实质收益。 Qwen-VLA-Base 已达到合理性能（LIBERO 90.8%, Simpler-WidowX 64.3%），表明大规模预训练学到跨本体的可转移操作原始体。通过指令调优，Qwen-VLA-Instruct 在所有基准上持续改进（LIBERO +7.1%，RoboCasa-GR1 +16.3%，Simpler-WidowX +9.4%，RoboTwin-Easy +21.8%，RoboTwin-Hard +20.8%），适量的任务特定对齐足以将一般预训练表示转换为精确、可部署的策略。

5.1.2 真实世界中的操作结果

我们在 ALOHA 上收集真实世界数据集，一个代表性的双臂机器人本体。该平台由两个 6-DoF 机械臂组成，配备平行夹爪，形成双臂操作系统。配备三个 RGB 摄像机：两个腕部安装摄像机和一个第一人称视图摄像机。与在仿真环境上微调的 Qwen-VLA-Instruct 不同，我们通过在 ALOHA 收集的多个任务演示上微调来验证 Qwen-VLA 在真实世界任务上的性能。我们准备两个共享相同架构的变体：Qwen-VLA-aloha$_{\text{w/o pretrain}}$ 从头开始训练，Qwen-VLA-aloha$_{\text{w/ pretrain}}$ 从 Qwen-VLA-Base 微调，以评估大规模机器人化预训练到真实世界设置的可转移性。

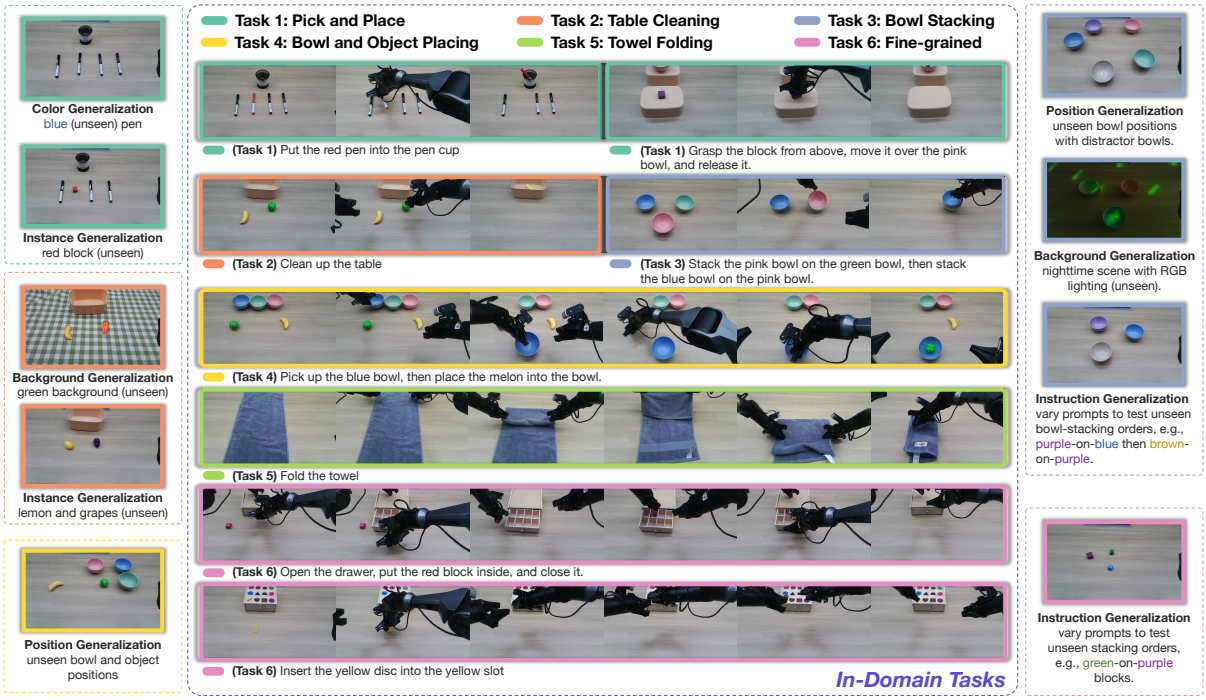

图 4: ALOHA 双臂平台上真实世界评估任务的概览。

我们在多样化的真实世界操作任务集上评估模型，涵盖分布内和分布外（OOD）场景。分布内评估包含六个任务类别：拾取和放置、桌面清洁、碗堆叠、碗拾取和物体放置、毛巾折叠、细粒度操作。OOD 评估测试对无法看到的颜色、实例、位置、背景、语言指令的泛化。

分布内任务。 分布内评估涵盖六个复杂度递增的任务类别：

- **任务 1：拾取和放置**。机器人拾起基础物体，包括积木和笔，并将其放在指定的目标位置。
- **任务 2：桌面清洁**。机器人清理表中的物体并将其放入目标容器。
- **任务 3：碗堆叠**。机器人按语言指令指定的颜色顺序堆叠碗。
- **任务 4：碗拾取和放置**。机器人拾起一个碗，根据语言指令将目标物体放入其中。
- **任务 5：毛巾折叠**。机器人在桌子上折叠毛巾，测试其操作可变形物体的能力。
- **任务 6：细粒度操作**。此类别包括多个精确操作任务，例如打开抽屉放置物体，然后关闭它，将物体插入狭缝和堆叠两个积木。

OOD 任务。 OOD 评估包括五个评估不同类型泛化的设置：

- **Color Generalization.** This setting evaluates color generalization on Task 1 and Task 4, where the robot needs to handle unseen color specifications, such as picking up a bowl with a specified color (unseen) and placing an object into it, or putting a pen with a specified color (unseen) into the target container.
- **Instance Generalization.** This setting evaluates instance generalization on Task 1 and Task 2, where the robot encounters novel objects, such as grapes (unseen) and a lemon (unseen) during table cleaning, or a red block (unseen) during object placement.
- **Position Generalization.** This setting evaluates position generalization on Task 3 and Task 4, where bowls and target objects appear at spatial configurations (unseen).
- **Background Generalization.** This setting evaluates background and lighting generalization on Task 2

and Task 3, including lighting conditions (unseen) for bowl stacking and a green background (unseen) for table cleaning.

- **Instruction Generalization.** This setting evaluates instruction generalization on Task 3 and Task 6, where the robot is required to execute different actions according to the given language instruction. Examples include stacking bowls in a specified order (unseen) or stacking blocks according to color relations, e.g., placing the green block (unseen) on the purple block (unseen).

表 5: **In-domain performance across short-horizon and long-horizon task categories.** Qwen-VLA-aloha$_{\text{w/o pretrain}}$ and Qwen-VLA-aloha$_{\text{w/ pretrain}}$ share the same Qwen-VLA model architecture; the former is trained from scratch, and the latter is fine-tuned from Qwen-VLA-Base.

Model	Short-horizon Tasks			Long-horizon Tasks			Avg.
	Pick and Place	Table Cleaning	Bowl Stacking	Bowl Pick & Place	Towel Folding	Fine-grained Manipulation	
GR00T N1.6 (NVIDIA et al., 2025)	30.8	38.5	53.8	19.2	19.2	10.3	28.6
$\pi_{0.5}$ (Black et al., 2025)	73.1	84.6	88.5	69.2	**80.8**	33.3	71.6
Qwen-VLA-aloha$_{\text{w/o pretrain}}$	30.8	53.8	61.5	64.1	50.0	30.8	48.5
Qwen-VLA-aloha$_{\text{w/ pretrain}}$	**96.2**	**92.3**	**98.7**	**87.2**	65.4	**61.5**	**83.6**

分布内任务结果。 表 5 总结分布内结果。Qwen-VLA-aloha$_{\text{w/ pretrain}}$ 达到最佳平均成功率 83.6%，在大多数任务类别上超越强基线，包括 GR00T N1.6 (NVIDIA et al., 2025) 和 $\pi_{0.5}$ (Black et al., 2025)。大规模预训练的益处明显：从 Qwen-VLA-Base 微调将平均成功率从 48.5% 提升到 83.6%。在拾取和放置、桌面清洁、碗堆叠、碗拾取和放置、细粒度操作上获得特别大的改进。这些结果确认预训练为真实世界操作提供强大基础，有效转移到多样化分布内任务。

OOD 任务结果。 表 6 报告 OOD 泛化性能。Qwen-VLA-aloha$_{\text{w/ pretrain}}$ 在所有五个泛化设置中持续达到最佳结果：颜色、实例、位置、背景和指令泛化。其平均 OOD 成功率达 76.9%，比 $\pi_{0.5}$ (Black et al., 2025) 超越 35.4 个百分点，比 Qwen-VLA-aloha$_{\text{w/o pretrain}}$ 超越 40.7 个百分点。改进在背景和指令泛化下尤其显著，分别达 80.8% 和 84.6%。这些结果表明预训练不仅改进分布内性能，还提供对看不到的视觉条件、物体实例、位置和指令变异的更强稳健性。

总体分析。 结果表明预训练在真实世界操作学习中起关键作用。尽管两个变体共享相同架构，从头开始训练导致性能低很多，特别是在 OOD 设置下。性能增益不单纯来自模型架构，而主要来自预训练的 Qwen-VLA-Base 模型。此外，Qwen-VLA-aloha$_{\text{w/ pretrain}}$ 的强大 OOD 性能表明预训练视觉-语言-动作表示能有效支持跨视觉变异和指令变异的泛化。

5.1.3 导航结果

我们在视觉和语言连续环境中导航（VLN-CE (Krantz et al., 2020; Ku et al., 2020)）上评估 Qwen-VLA 的导航能力。我们在 R2R 和 RxR 基准的 Val-Unseen 分割上测试 Qwen-VLA Base 和 Instruct 版本，与开源基线进行比较。我们使用 VLN-CE 的默认设置，特别是实现与 Qwen-VLA 预测轨迹兼容的 `sliding-window` 路径点动作。

表 7 显示，Qwen-VLA-Instruct 在 R2R 和 RxR 基准上的大多数指标上达到最佳性能。在 R2R Val-Unseen 上达到最高预言成功率（69.0）和成功率（57.5），分别超越 StreamVLN 4.8 和 0.6 个百分点，同时保持竞争 SPL。在更具挑战性的 RxR Val-Unseen 分割上，在 SR（59.6）和 SPL（47.8）上领先，显著超越所有基线。VLA 和 VLN 数据上的联合训练在两个基准上保持合理性能。

5.1.4 静态操作的分布外评估

我们构造控制 OOD 基准来测试预训练先验是否泛化超出微调分布：微调仅涵盖简单拾取-放置但评估需要看不到的任务类型。我们开发 SimplerEnv-OOD，一套 6 个分布外任务跨越 3 个桌面场景，在 SimplerEnv 框架基础上用 WidowX 机器人臂。所有模型仅在 Bridge 训练分割上微调，仅包含具有固定物体配对的简单拾取-放置演示。没有任何 OOD 任务指令、空间关系或操作原始体出现在训练数据中。六个任务为：

表 6: **OOD performance across generalization categories.** Qwen-VLA-aloha$_{\text{w/o pretrain}}$ and Qwen-VLA-aloha$_{\text{w/ pretrain}}$ share the same Qwen-VLA model architecture; the former is trained from scratch, and the latter is fine-tuned from pretrained Qwen-VLA-Base.

Model	Color	Instance	Position	Background	Instruction	Avg.
GR00T N1.6 (NVIDIA et al., 2025)	46.2	38.5	3.8	19.2	19.2	25.4
$\pi_{0.5}$ (Black et al., 2025)	57.7	61.5	19.2	26.9	42.3	41.5
Qwen-VLA-aloha$_{\text{w/o pretrain}}$	42.3	30.8	34.6	30.8	42.3	36.2
Qwen-VLA-aloha$_{\text{w/ pretrain}}$	**88.5**	**76.9**	**53.8**	**80.8**	**84.6**	**76.9**

表 7: **与 VLN-CE 上开源基线的比较**。我们将 Qwen-VLA Base 和 Instruct 版本与广泛认可的开源基线在 R2R 和 RxR 基准的 Val-Unseen 分割上进行比较。**粗体**：最佳；下划线：第二最佳。

Method	R2R Val-Unseen				RxR Val-Unseen			
	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
NaVid (Zhang et al., 2024)	5.7	49.2	41.9	36.5	**5.7**	45.7	38.2	–
Uni-NaVid (Zhang et al., 2025b)	5.6	53.3	47.0	42.7	6.2	48.7	40.9	–
NaVILA (Cheng et al., 2025)	5.2	62.5	54.0	49.0	6.8	49.3	44.0	58.8
StreamVLN (Wei et al., 2025)	**5.0**	64.2	56.9	**51.9**	6.2	52.9	46.0	**61.9**
Qwen-VLA-Base	5.2	61.7	53.8	49.4	6.4	55.1	45.8	56.2
Qwen-VLA-Instruct	5.1	**69.0**	**57.5**	51.2	5.8	**59.6**	**47.8**	57.1

- **MoveAway：**将勺子移离布料更远，使最终距离超过初始距离。这通过距离增加放置测试位置泛化。
- **MoveRight：**将勺子移到布料的右侧。这通过具有正交轴约束的方向放置测试位置泛化。
- **PlaceNear：**将胡萝卜移到相邻但不在盘子顶部的位置。这通过基于接近度的放置测试位置泛化。
- **PlaceRight：**将勺子放在布料的右半区域。这通过区域约束放置测试位置泛化，需要策略解释子区域说明符（“右半”）并在指定区域内实现正确目标接触和空间精度。
- **PutFront：**将胡萝卜放在盘子前面。这沿着前/后方向轴测试位置泛化。
- **StackYellow：**拾起黄色积木并将其堆叠在绿色积木上。训练数据仅包含在黄色上堆叠绿色；这个颠倒顺序通过需要基于颜色的推理和泛化到在训练期间未见过的新颖物体-颜色绑定测试视觉泛化。

表 8 显示，Qwen-VLA-Instruct 达到最高平均成功率 **32.0%**，在大多数任务上显著超越 $\pi_{0.5}$（12.6%）。优势在位置泛化上最强：$\pi_{0.5}$ 在 MoveRight 和 PlaceNear 上完全失败，Qwen-VLA-Instruct 分别达 33.3% 和 39.6%。一致增益出现在 MoveAway（43.8% vs. 26.1%）和 PlaceRight（47.9% vs. 32.1%）上，表示对看不到的空间指令更强的泛化。在 StackYellow（22.9% vs. 4.2%）上，其中训练数据仅包含在黄色上堆叠绿色但任务需要颠倒顺序，Qwen-VLA-Instruct 表现出对新颖物体-颜色绑定更强的泛化。

表 8: **SimplerEnv-OOD 上的 OOD 泛化结果**。**粗体**：最佳；下划线：第二最佳。

Method	OOD Task						Avg.
	MoveAway	MoveRight	PlaceNear	PlaceRight	PutFront	StackYellow	
$\pi_{0.5}$ (Black et al., 2025)	26.1	0.0	0.0	32.1	**13.0**	4.2	12.6
Qwen-VLA-Base	31.3	31.6	16.7	47.1	6.3	18.8	25.3
Qwen-VLA-Instruct	**43.8**	**33.3**	**39.6**	**47.9**	4.2	**22.9**	**32.0**

5.1.5 动态操作的分布外评估

在复杂环境中部署机器人需要掌握动态操作并适应独立物体运动 (Liang et al., 2026)。DOMINO 基准 (Fang et al., 2026) 为系统地评估这个关键能力开创综合评估范例。它创新性地引入分层运动设计和连续操作分数来在不可预测约束下捕获执行质量。这些独特特性使 DOMINO 成为异常压力测试，确定视觉-语言-动作模型是否拥有可泛化空间到运动学映射能力，而不仅仅执行记忆的静态例程。我们在零样本设置中跨所有 35 个 DOMINO 套件评估 Qwen-VLA，报告成功率（SR）和操作分数（MS）。

表 9: **与 DOMINO 上最先进方法的比较。粗体**：最佳；下划线：第二最佳。

Method	SR (%)↑	MS↑
Fine-tuned on dynamic manipulation data		
OpenVLA (Kim et al., 2024)	1.5	6.1
RDT-1B (Liu et al., 2025b)	5.3	17.7
π_0 (Black et al., 2024)	8.2	24.0
$\pi_{0.5}$ (Black et al., 2025)	9.6	26.2
InternVLA-M1 (Chen et al., 2025b)	5.4	27.6
VLA-Adapter (Wang et al., 2026)	4.4	24.3
π_0-FAST (Pertsch et al., 2025)	3.5	20.9
OpenVLA-OFT (Kim et al., 2025)	9.1	24.1
StarVLA-OFT (Community, 2026)	10.9	30.5
PUMA (Fang et al., 2026)	17.2	35.0
Zero-shot to dynamic manipulation		
OpenVLA-OFT (Kim et al., 2025)	6.7	20.0
$\pi_{0.5}$ (Black et al., 2025)	7.5	20.4
LingBot-VLA w/ depth (Wu et al., 2026)	11.8	26.7
LingBot-VA (Li et al., 2026a)	<u>24.1</u>	36.1
Qwen-VLA-Base	21.1	<u>37.4</u>
Qwen-VLA-Instruct	**26.6**	**39.5**

表 9 显示，Qwen-VLA 在 DOMINO 基准上建立最佳总体性能。Qwen-VLA-Instruct 达到最高 SR（26.6%）和 MS（39.5）。在零样本类别中，我们的模型在 SR 中超越 OpenVLA-OFT (Kim et al., 2025) 和 $\pi_{0.5}$ (Black et al., 2025) 超过 19 个百分点，证明相对标准视觉-语言-动作架构的重大优势。在动态对象动力学下超越代表性 WAM 风格方法 LingBot-VA (Li et al., 2026a)，展现更强连续执行质量。至关重要的是，Qwen-VLA 超越明确为动态操作微调的整个基线套件。高级方法 PUMA (Fang et al., 2026) 依赖 DOMINO 特定微调和时间运动输入来维持动态感知。尽管缺乏这些定制适应并仅依赖当前帧观测，Qwen-VLA-Instruct 在 SR 中超越 PUMA 9.4 个百分点，MS 中 4.5 点。我们的模型在推理时仅使用当前帧观测达到这些收益，无需任何动态操作微调。这突显统一动作和轨迹预训练策略在学习可转移空间到运动学先验中的优势。

这个分布外泛化归因于决定性动作生成和广泛机器人化预训练的组合。流匹配动作解码器产生连贯动作块，减少犹豫并帮助策略在狭窄时间窗口内精确作用。跨操作、导航、轨迹预测和视觉-语言数据的大规模联合预训练提供可转移的视觉接地、空间推理和连续控制先验。这些稳健先验使 Qwen-VLA 能有效泛化到移动物体操作，尽管完全缺乏动态训练数据。

5.1.6 真实世界中的分布外泛化

为验证连续预训练，我们在四种分布外设置下评估 **Qwen-VLA-Base** 在 ALOHA 双臂机器人上的性能：颜色定位、新物体操作、未见物体识别和背景鲁棒性。代表性展示见 Figure 5。

颜色定位（左上）。 我们测试 Qwen-VLA-Base 是否能区分仅颜色不同的视觉相似物体。在四次试验中，机器人被提示" 抓住 {color} 球"，分别对应绿色、蓝色、红色和黄色。在所有情况下，模型正确识别了目标颜色并生成了准确的到达-抓取轨迹，表明经过连续预训练后的颜色条件化语义定位。

新物体抓取和组合清理（右上）。 我们测试模型是否能零样本抓取操作预训练数据中从未见过的物体。前两面板展示成功抓取：由" 抓住 {object}" 提示的绿色西兰花和玩具鸭，都不在操作训练集中。模型成功定位并抓住西兰花；对玩具鸭，模型正确识别并靠近，尽管不规则形状阻止稳定抓取。后两面板展示由" 清理桌子" 提示的组合任务：机器人依次拿起蓝色雨伞、玩具鸭和酸奶瓶放入箱子，这三者都不在操作集中。Qwen-VLA-Base 顺序完成清理，展现对新物体的零样本视觉识别和拿取-放置循环后的重新规划能力。这与混合 VL–VLA–VLN 训练一致：视觉-语言数据提供丰富物体词汇表，使模型即使无域内演示也能迁移到操作任务。

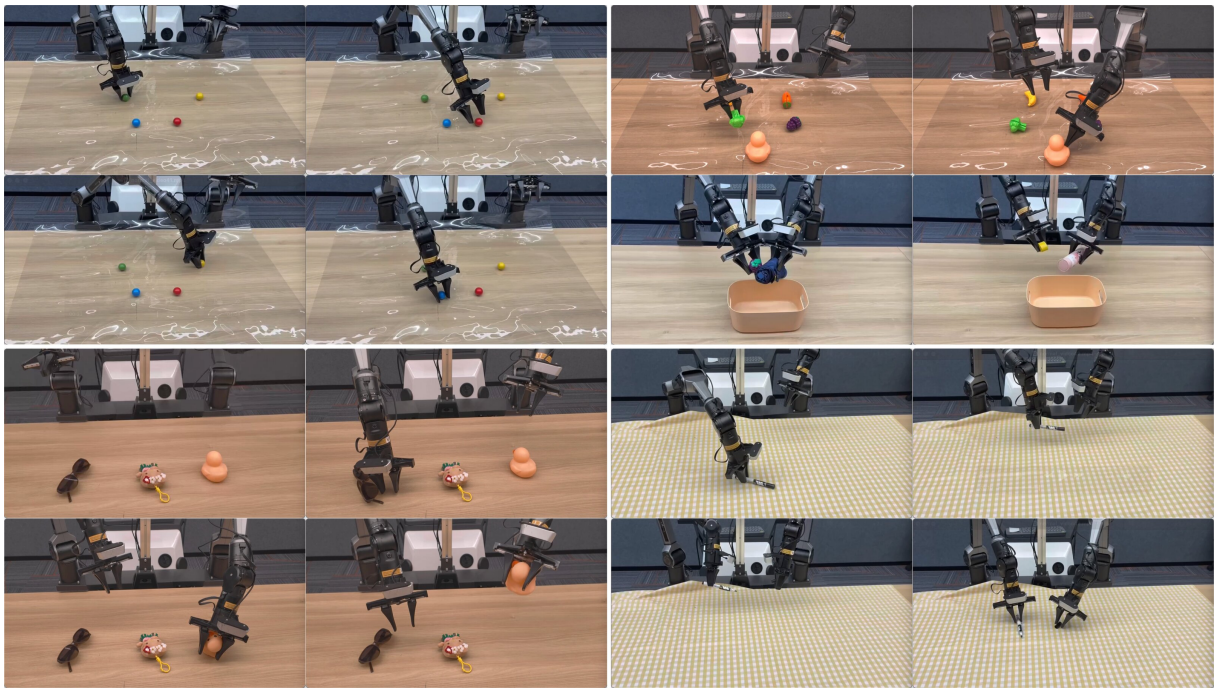

图 5: **Qwen-VLA-Base 在 ALOHA 双臂机器人上的定性分布外泛化。左上**：绿色、蓝色、红色和黄色球的颜色条件化抓取。**右上**：上两个面板显示新物体（绿色西兰花、玩具鸭）的抓取；下两个面板展示一个组合式" 清理桌子" 任务，涉及序列化的拿取并放入箱子（蓝色雨伞、玩具鸭、酸奶瓶）。**左下**：与完全未见过的物体互动（太阳镜、毛绒娃娃、玩具鸭）。**右下**：原子动作迁移通过打开笔帽并放置笔帽，对未见过的黄色背景具有鲁棒性。

未见物体交互（左下）。 我们进一步探测零样本泛化，对预训练中完全缺失的物体类别发出" 接近{object}" 指令：太阳镜、毛绒娃娃和玩具鸭。" 接近" 动作在操作数据中几乎不存在，但模型正确解释指令并将末端执行器导向目标物体。对玩具鸭，模型甚至通过抓住头部成功抓取。对太阳镜，模型准确定位并接近，尽管细长平坦几何形状阻止稳定抓取。这反映形状泛化的物理边界，而非视觉识别或指令跟随失败。结果显示混合 VL–VLA–VLN 连续预训练的指令跟随迁移能力：模型跟随新颖动作动词，利用通用预训练数据的语言和视觉覆盖识别未见物体。

背景鲁棒性（右下）。 最后，我们探测对视觉域转移的鲁棒性。Qwen-VLA-Base 成功打开笔帽并放在桌子上，这需要精确力量控制的灵巧两阶段动作。虽然任务在预训练期间可能已见过，但黄色背景在训练中未见。策略尽管面对未见视觉背景仍完成任务且性能无下降。背景扰动不变性与连续预训练的通用 VL 理解数据视觉多样性一致。模型接触多样场景、照明和背景的自然图像，促使视觉编码器关注任务相关物体而非虚假背景线索。

5.2 消融实验

5.2.1 文本到动作预训练消融实验

如 Section 3.1 所述，T2A 在视觉接地前，让解码器仅从语言和本体感知提示中学习重建动作分布，建立结构化的动作先验。我们对塑造这一先验的五个设计选择进行消融：数据组成、序列预测模式、视觉输入、流匹配时步分布和训练时长。所有消融实验共享相同的下游训练策略，结果为在 Simpler-WidowX 上从 T2A 检查点微调后的任务成功率（%）。

数据组成。 T2A 仅从语言获取动作先验，故其数据组成直接决定了解码器的学习内容。我们考虑两个数据源。*Syn*（仿真流程生成的纯语言轨迹）具有低成本扩展和广泛的任务覆盖，但轨迹在运动学上理想化了。*Real*（真机遥操作数据，去除视觉）捕捉真实物理动力学，但收集成本高、任务覆盖窄。

如图 6（a）所示，纯真实数据得 51.0%，纯综合数据得 64.1%——虽有改进，但仍非最优。混合 20%

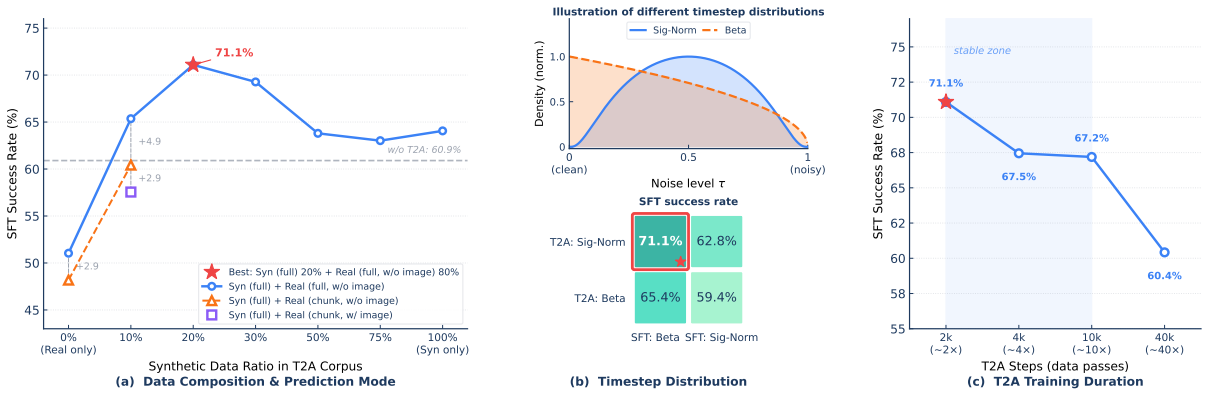

图 6: T2A 预训练消融实验。**(a) 数据组成和预测模式。** T2A 语料库中的微调成功率（%）对比综合数据比例。虚线标记无 T2A 的基线（60.9%）。全序列预测与约 20% 的综合数据 + 80% 的真实数据达到最佳结果（71.1%，相比无 T2A 提升 10.2 个百分点）。分块预测在全序列中表现一致较差（例如在 10% 综合数据时提升 4.9 个百分点），在 T2A 期间包含图像 token 进一步降低了分块模式性能（下降 2.9 个百分点），证实应在 T2A 中完全抑制图像。**(b) 流匹配时步分布。** T2A 中的 Sigmoid-Normal $p(\tau)$ 与微调中的 Beta $p(\tau)$ 的组合达到最高成功率（71.1%）；反转任一阶段的选择都会降低性能。**(c) T2A 训练时长。**性能在 2000 步时达到峰值（71.1%）并在 10000 步时保持平台；40000 步由于对 T2A 语料库过拟合而略有下降。

综合与 80% 真实数据达最佳（**71.1%**），较无 T2A 基线（60.9%）提升 10.2 个百分点。综合数据扩大语言-动作映射的覆盖范围，真实数据则将先验约束在物理合理的动力学。

序列预测模式：全长度对比分块。 全序列预测在单个前向传播中监督整个轨迹，让解码器学习长视野时间结构和自然的起始/终止模式。分块预测则将轨迹分割成固定长度窗口，训练解码器仅从指令生成下一个窗口，这会截断长视野依赖和在块边界处引入位置歧义。

图 6（a）显示全序列预测在各数据比例下都优于分块预测。10% 综合数据时差 4.9 个百分点（65.4% vs 60.4%），纯真实数据时差 2.9 个百分点（51.0% vs 48.2%）。完整轨迹让解码器学到语言与完整动作序列的映射关系（包含轨迹级连贯性和组合性），而分块仅提供局部片段，映射不完整。动作表示在 T2A 和下游阶段保持一致：编码为每个分块第一帧相对的末端执行器增量位移，本体感知提示指定机器人平台、归一化约定和预测视野，使解码器能跨异质本体在相同目标内工作。

T2A 期间的视觉输入。 T2A 刻意设计为无视觉阶段。包含图像观测会增加计算成本约一个数量级，更关键的是会让解码器依赖视觉线索，而非学习语言-动作接地。

图 6（a）验证了这点：10% 综合数据下，含图像的分块预测得 57.6%，去图像的相同设置达 60.4%——罚分 2.9 个百分点。有图像时，解码器可偷懒依赖视觉相关性，而不构建可靠的语言-动作映射，稀释了 T2A 本应建立的先验。故应在 T2A 期间完全抑制图像 token，将视觉接地延迟到 CPT 阶段，此时解码器已掌握良好的动作先验。

流匹配时步分布。 流匹配策略条件化于标量时步 $\tau \in [0,1]$，在清晰目标（τ=0）和纯噪声（τ=1）间插值。抽样分布 $p(\tau)$ 决定训练中哪些噪声水平接收最多梯度。标准做法用 *Beta* 分布（密度集中在清晰端），我们在 CPT 和 SFT 阶段遵循此约定。但 T2A 在无视觉条件下运作，解码器无骨干特征指导去噪，信号到噪声比最优的地方在中间时步。因此我们比较 Beta 和 *Sigmoid-Normal* 分布（后者在中间噪声处达峰值，图 6（b））。

如图 6（b）所示，T2A 用 Sigmoid-Normal、微调用 Beta 的组合达最高成功率（**71.1%**）。T2A 中用 Beta 降至 65.4%（-5.7pp），微调中用 Sigmoid-Normal 降至 62.8%（-8.3pp），两阶段都用 Beta 最差（59.4%）。在无视觉条件指导下，中间时步对语言-动作先验的学习信号最强；一旦视觉-语言骨干在 CPT/微调提供丰富条件化，梯度更均匀的 Beta 形状 $p(\tau)$ 样本效率更高。故最终流程中 T2A 用 Sigmoid-Normal，CPT/微调用 Beta。

T2A 训练时长。 更长的 T2A 训练让解码器更充分探索动作分布，但在固定语料库上过度训练易过拟合其特异性，减少对 CPT 的可塑性。我们扫描广泛的 T2A 时长找最优点。

如图 6（c）所示，性能在 2000 步达峰值（**71.1%**），4000 步（67.5%）和 10000 步（67.2%）保持相近，40000 步显著降至 60.4%。快速收敛在预期内：T2A 学的语言-动作映射在结构上维度低于完整视觉-语言-动作空间，少量步骤足以捕捉；延长训练反而开始记忆特定轨迹实例而非细化结构先验。故最终流程中采用 **2000 T2A 步**作为默认值，成本-质量权衡最优。

5.2.2 多本体联合训练

我们研究两个对多本体联合训练至关重要的设计轴：动作学习期间视觉-语言数据的角色和异构动作空间的投影策略。

视觉-语言数据对动作学习的影响。 虽然视觉-语言监督在离散令牌空间中运作，动作预测以连续轨迹为目标，两个目标共享通用感知和语言主干。自然问题是在机器人化微调期间保留 VL 数据是否有助于或伤害动作质量。我们在图 7 中比较三个配置：**(a)** *仅动作*：主干仅在机器人化轨迹上微调，无 VL 监督；**(b)** *动作 +VL 数据*：VL 数据的一部分（VQA、标题、空间接地）混合到训练语料库。

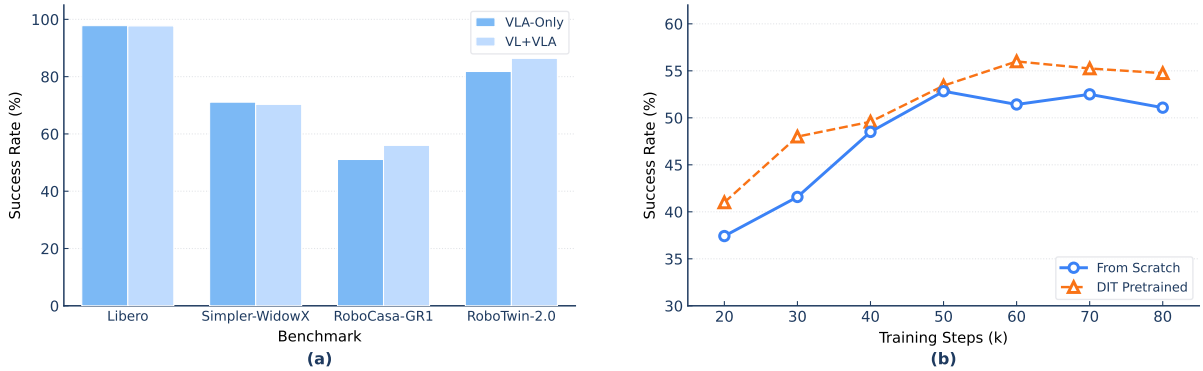

图 7: 视觉-语言联合训练消融。**(a) VL 数据对动作学习的影响。**跨四个基准的任务平均成功率 (%) 为 VLA-Only （仅动作数据）vs. VL+VLA（动作 + 视觉-语言数据联合训练）。在更简单基准（Libero、Simpler-WidowX）上两个配置性能相当，确认 VL 联合训练不引入干扰。在需要细粒度物体识别和组合指令解析的基准上，VL+VLA 产生明显收益：RoboCasa-GR1 +4.9 pp (51.1% → 56.0%) 和 RoboTwin-2.0 +4.6 pp (81.8% → 86.4%)。**(b) 预训练 DiT 的可转移性。** RoboCasa-GR1 上 SFT 成功率 (%) vs. 训练步骤 (k) 当将原始 Qwen3.5 VLM 主干与从零初始化的 DiT 解码器 vs. 预训练 DiT 配对时。预训练 DiT 收敛更快并达更高峰值。

图 7 (a) 显示，将 VL 数据混合到训练中在操作基准上产生明确益处。在 Libero 和 Simpler-WidowX 上，VLA-Only 和 VL+VLA 达到近似相同的成功率。在 RoboCasa-GR1 上，这需要在混乱厨房场景中具有多样物体和空间配置的全身控制，VL+VLA 比 VLA-Only 改进 **+4.9%** 绝对值（51.1% → 56.0%）。此外，我们将预训练 DiT 附加到新鲜 Qwen3.5-4B VLM 并与从零开始基线比较（DiT 随机初始化）。图 7 (b) 显示，预训练 DiT 在整个训练中持续超越从零开始基线，在早期阶段收敛更快并达更高峰值成功。

异构本体的投影设计。 不同机器人平台有不同动作维度和语义（例如 7-DoF 臂 vs. 29-DoF 全身）。为了跨异构动作格式学习共享表示空间，DiT 隐空间和按本体动作向量之间的投影必须协调这个不匹配。为此，我们比较三个代表性设计。

多-MLP。每个本体拥有私有编码器 MLP，将其本机动作维度映射到 DiT 隐大小，和私有解码器 MLP 映射回来。令 d_i 表示本体 i 的动作维度，h DiT 隐大小，N 本体数。每个本体拥有私有编码器 $f_i^{\text{enc}}: \mathbb{R}^{d_i} \to \mathbb{R}^h$ 和解码器 $f_i^{\text{dec}}: \mathbb{R}^h \to \mathbb{R}^{d_i}$，总共 $2h\sum_{i=1}^{N} d_i$ 参数。

连接。来自所有 N 个本体的动作连接到单个维度 $D = \sum_{i=1}^{N} d_i$ 的向量；每个本体占有专用、不重叠段。共享编码器 $f^{\text{enc}}: \mathbb{R}^D \to \mathbb{R}^h$ 和解码器 $f^{\text{dec}}: \mathbb{R}^h \to \mathbb{R}^D$ 处理完整向量（$2h\sum_{i=1}^{N} d_i$ 参数），每个本体从对应段检索其预测。

零填充。单个共享 MLP 编码所有本体；较短动作向量右填充零到 $d_{\max} = \max_i d_i$ 在编码前。共享编码器 $f^{\text{enc}}: \mathbb{R}^{d_{\max}} \to \mathbb{R}^h$ 和解码器 $f^{\text{dec}}: \mathbb{R}^h \to \mathbb{R}^{d_{\max}}$ 处理填充向量（$2h\,d_{\max}$ 参数）；每个本体仅读其 d_i 维前缀。由于 $d_{\max} \le \sum_i d_i$，*零填充*使用比*多-MLP* 和*连接*更少投影参数。

表 10: **投影设计消融**。结果是任务平均成功率（%）。所有联合训练变体共享相同训练方案。

	Single-Embodiment Training		Multi-Embodiment Co-training		
	Bridge Only	Robocasa Only	Multi-MLP	Concat.	Zero-Pad
Bridge	62.8	—	**63.3**	63.0	63.0
Robocasa	—	53.4	52.1	52.8	**53.2**

表 10 显示，所有联合训练投影设计在两个基准上达到与单本体基线相当的性能（Bridge 单训 62.8% vs. 联合训 ∼63%；Robocasa 单训 53.4% vs. 联合训 52–53%），多本体联合训练相对于在每个数据集上独立训练无显著惩罚。在三个设计中，性能差异微小（两个基准上 <1.2% 绝对值），一旦建立共享隐空间，投影架构选择对任务成功的影响有限。然而零填充在架构上最轻：仅需 $2h\,d_{\max}$ 投影参数，对比多-MLP 和连接的 $2h\sum_i d_i$。因此我们采用 **零填充**作为默认投影设计。

5.2.3 强化学习后训练的效果

我们研究 SFT 之上的 RL 是否提供额外收益，以及这些收益是否转移到 RL 训练期间看不到的基准。为隔离每个后训练阶段的贡献，我们评估沿训练管道的三个检查点：CPT 基础模型（**Qwen-VLA-Base**）、多任务 SFT 后的模型（**+SFT**）和 RL 精细化后的模型（**+RL**，即 Qwen-VLA-Instruct）。RL 展开仅在 SimplerEnv 中收集，具有稀疏二元成功奖励；RL 阶段期间不使用其他基准数据。

表 11: **后训练阶段的累积效果**。每行在前一个检查点基础上添加一个阶段。RL 展开仅在 SimplerEnv 中收集。除 DOMINO 操作分数（MS）外所有数字都是成功率（%）。

Stage	**In-Distribution**					**OOD-static**	**OOD-dynamic (DOMINO)**	
	Simpler	RoboCasa	RoboTwin-E	RoboTwin-H	LIBERO	SimplerOOD	SR	MS
CPT	64.3	40.4	64.3	66.4	90.8	25.3	21.1	37.4
+ SFT	70.8	56.0	86.3	87.1	97.8	31.6	25.7	39.1
+ RL	**73.7**	**56.7**	86.1	**87.2**	**97.9**	**32.0**	**26.6**	**39.5**

表 11 显示，每个后训练阶段提供补充收益。SFT 通过将通用 CPT 表示专门化为下游任务分布来跨所有基准传递大型改进。RL 在 SFT 之上提供进一步上升，在 SimplerEnv 上最大收益（+2.9 pp，从 70.8% 到 73.7%），即 RL 展开收集的环境。这确认直接优化闭环任务成功捕获模仿目标无法单独捕获的策略改进。

更重要的是，RL 收益不限于训练环境。在不属于 RL 展开分布的基准上，性能保留或温和改进：RoboCasa 增加 0.7 pp, RoboTwin-Hard 0.1 pp, LIBERO 0.1 pp, SimplerEnv-OOD 上 OOD 评估 0.4 pp。RoboTwin-Easy 表现出可忽略变化（<0.6 pp），RL 不对保留任务导致灾难遗忘。在 DOMINO 动态操作基准上（测试对移动物体的零样本泛化），尽管 RL 训练期间完全缺乏动态操作数据，SR（25.7% → 26.6%）和 MS（39.1 → 39.5）都改进。这些结果表明仿真中的任务成功优化保留性能具有温和正转移：策略学到更坚定执行和从复合错误恢复，跨评估设置泛化的益处。

5.2.4 状态条件化

许多机器人学习方法在视觉观测之外将策略条件化为显式本体感知状态（例如关节角），基于已知当前机器人配置帮助预测下一个动作的假设。然而本体感知输入也引入可能阻碍跨本体泛化的本体特定依赖性，因为不同平台有不同关节空间和运动学结构。我们检验显式状态条件化是否在 Qwen-VLA 中惠益动作学习，如果是，状态应在体系结构中何处注入。我们在 RoboTwin-2.0 上比较三个策略：

- **无状态**。模型仅接收视觉观测和语言指令。没有本体感知信息提供给 VLM 主干或 DiT 动作解码器。

- **VLM 提示中的状态**。关节角离散化为每维 256 个箱并序列化到语言提示，例如 `"state: [34, 128, 67, 200, 55, 12, 88] pick up the bowl"`。VLM 将状态令牌与图像和指令一起处理；DiT 收不到额外状态输入。
- **DiT 中的状态**。关节角作为连续值向量直接传递给 DiT 动作解码器作为额外条件化输入，完全绕过 VLM。

表 12: RoboTwin-2.0 上的状态条件化消融（成功率，%）。

Conditioning	RoboTwin-Easy	RoboTwin-Hard
No State	88.7	87.4
State in VLM Prompt	89.3	88.7
State in DiT	89.4	88.3

表 12 显示，包含本体感知状态仅产生边际收益，无论注入策略如何：相对于仅视觉基线在 Easy 上最多 +0.7 pp，Hard 上 +1.3 pp。两个注入策略性能相当（$\leq$0.4 pp 差异），VLM 的语义处理和 DiT 的连续条件化都不为编码关节角信息提供明确优势。状态条件化的有限影响归因于两个因素。首先，多视图视觉观测已提供关于机器人当前配置的充足信息，特别是对于末端执行器在至少一个摄像机视图中可见的任务。其次，流匹配动作解码器预测相对动作位移而非绝对姿态，减少对显式当前状态参考的需要。考虑到有限的益处和维护跨多样化平台的本体特定状态接口的额外复杂性，我们选择在默认框架中不包含状态条件化，保持本体感知文本提示作为唯一平台特定输入。

6 结论

我们呈现 **Qwen-VLA**, 统一视觉-语言-动作模型, 将 Qwen 视觉-语言主干从感知和推理扩展到机器人化动作生成。通过在共享动作和轨迹预测框架内表述操作、导航、自我中心动作建模和轨迹预测，Qwen-VLA 使单个模型从跨任务、环境和机器人本体的异构机器人化数据学习。通过本体感知提示条件化、基于 DiT 的流匹配动作解码器、大规模联合预训练和 SFT/RL 后训练，Qwen-VLA 实现强大通用性能，同时保留不同控制约定的灵活性。我们的结果建议语言接地 VLA 模型通过对齐多模态感知、指令理解、本体约束和可执行动作提供朝向机器人化智能的实践路径。与视觉预测中心世界模型相比，Qwen-VLA 强调将多模态理解接地到可由机器人化智能体执行的动作中。这些结果表明 VLA 模型可作为多模态基础模型和机器人化智能体之间的可执行接口，桥接高级语言推理与低级物理控制。

7 局限性与未来工作

尽管具有通用能力，Qwen-VLA 仍有多个局限。首先机器人化动作数据仍比视觉-语言预训练数据小得多且多样性少，限制对长尾物体、环境、本体和接触丰富交互的稳健性。其次跨视觉-语言理解、导航和动作生成的联合训练引入优化权衡。虽然面向动作训练改进策略学习，它可温和回退一些纯视觉-语言和导航评估，需要更好目标平衡、数据课程和模块专门化。第三当前评估仍主要是短期和基准驱动，长期、易失败真实世界部署留作开放挑战。

这些局限性激励多个未来方向。通过自主收集、仿真和仿真到真实转移扩展真实世界交互数据对改进稳健性至关重要。大规模人类视频从自我中心和第三人称视点可提供物理先验和超越机器人演示的时间抽象。未来模型应包含长期规划、情节记忆和世界建模，使智能体能跟踪状态、分解任务、从失败恢复和预期动作后果。最后更丰富物理反馈如力、触觉和本体感知信号，与大规模强化学习在仿真和真实世界中结合，可能进一步桥接语言接地机器人化理解和可靠控制之间的差距。

8 贡献与致谢

核心贡献者

Qiuyue Wang* Mingsheng Li* Jian Guan* Jinhui Ye Sicheng Xie
Yitao Liu Junhao Chen Zhixuan Liang Jie Zhang Xintong Hu
Xuhong Huang Pei Lin Junyang Lin Dayiheng Liu Shuai Bai†
Jingren Zhou

贡献者

Jiazhao Zhang Haoqi Yuan Gengze Zhou Hang Yin Ye Wang
Yiyang Huang Zixing Lei Wujian Peng Delin Chen Yingming Zheng
Jingyang Fan Xianwei Zhuang Xin Zhou Haoyang Li Anzhe Chen
Tong Zhang Xuejing Liu Yuchong Sun Ruizhe Chen Zhaohai Li
Chenxu Lü Zhibo Yang Tao Yu Xionghui Chen

致谢

我们感谢 Jun Huang、Jianlou Si、Chao Wang 和 Nana Jiang 对仿真环境和强化学习的强大支持。

*Equal contribution. †Corresponding author.

参考文献

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL https://arxiv.org/abs/2503.06669.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b.

Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking, 2023. URL https://arxiv.org/abs/2309.01918.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. π0.5: A vision-language-action flow model for general robot control. Technical report, Physical Intelligence, 2025.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. RT-1: Robotics transformer for real-world control at scale, 2022. URL https://arxiv.org/abs/2212.06817.

Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21819–21830, 2024.

Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7817–7826, 2025a.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025b.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems (RSS)*, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. In *Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Matei Ciocarlie, Corey Goldfeder, and Peter Allen. Dexterous grasping via eigengrasps: A low-dimensional approach to a high-complexity problem. In *Robotics: Science and systems manipulation workshop-sensing and adapting to the real world*, 2007.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Charles Corbiere, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. Retrieval-based interleaved visual chain-of-thought in real-world driving scenarios. *arXiv preprint arXiv:2501.04671*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 2088–2098, 2019.

Karim Elmaaroufi, Liheng Lai, Justin Svegliato, Yutong Bai, Sanjit A Seshia, and Matei Zaharia. Graid: Enhancing spatial reasoning of vlms through high-fidelity data generation. *arXiv preprint arXiv:2510.22118*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 12606–12633, 2024.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot, 2023. URL https://arxiv.org/abs/2307.00595.

Heng Fang, Shangru Li, Shuhan Wang, Xuanyang Xi, Dingkang Liang, and Xiang Bai. Towards generalizable robotic manipulation in dynamic environments. *arXiv preprint arXiv:2603.15620*, 2026.

Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22030–22040, 2024.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, and Xianpeng Lang. Driveaction: A benchmark for exploring human-like driving decisions in vla models. *arXiv preprint arXiv:2506.05667*, 2025.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026.

Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, et al. RoboMIND 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence, 2025. URL https://arxiv.org/abs/2512.24653.

Xintong Hu, Xuhong Huang, Jinyu Zhang, Yutong Yao, Yuchong Sun, Qiuyue Wang, Mingsheng Li, Sicheng Xie, Yitao Liu, Junhao Chen, Yixuan Chen, Yingming Zheng, Shuai Bai, and Tao Yu. Finevla: Fine-grained instruction alignment for steerable vision-language-action policies, 2026. URL https://arxiv.org/abs/2605.27284.

Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 930–938, 2024.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, et al. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning, 2022. URL https://arxiv.org/abs/2202.02005.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pp. 9902–9915. PMLR, 2022.

Xiaosong Jia, Yuqian Shao, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive-vl: Benchmarks for closed-loop autonomous driving with vision-language models. *arXiv preprint arXiv:2604.01259*, 2026.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, et al. Galaxea open-world dataset and G0 dual-system VLA model, 2025. URL https://arxiv.org/abs/2509.00576.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset, 2024. URL https://arxiv.org/abs/2403.12945.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026b.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, 2026c.

Dingkang Liang, Cheng Zhang, Xiaopeng Xu, Jianzhong Ju, Zhenbo Luo, and Xiang Bai. Cook and clean together: Teaching embodied agents for parallel task execution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18415–18424, 2026.

Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. Adaptdiffuser: diffusion models as adaptive self-evolving planners. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 20725–20745, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yifeng Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation, 2025a. URL https://arxiv.org/abs/2410.07864.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, 2025b.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pp. 252–269, 2024.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). In *European Conference on Computer Vision*, pp. 264–273. Springer, 2024.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne,

Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

Abby O' Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017.

Ropedia. Xperience-10m: A large-scale egocentric multimodal dataset of human experience. https://huggingface.co/datasets/ropedia-ai/xperience-10m, 2026.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, Nathan Ratliff, and Dieter Fox. curobo: Parallelized collision-free minimum-jerk robot motion generation, 2023. URL https://arxiv.org/abs/2310.17274.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy, 2025. URL https://arxiv.org/abs/2511.16651.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale, 2024. URL https://arxiv.org/abs/2308.12952.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024.

Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. TrackVLA: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the computer vision and pattern recognition conference*, pp. 22442–22452, 2025b.

Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 18638–18646, 2026.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, Shichao Fan, Xinhua Wang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Ning Liu, Pei Ren, Qiang Zhang, Yaoxu Lyu, Mengzhen Liu, He Jingyang, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Robotics: Science and Systems XXI*, RSS2025. Robotics: Science and Systems Foundation, June 2025a. doi: 10.15607/rss.2025.xxi.152. URL http://dx.doi.org/10.15607/RSS.2025.XXI.152.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, et al. RoboCOIN: An open-sourced bimanual robotic data collection for integrated manipulation, 2025b. URL https://arxiv.org/abs/2511.17441.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Shuailei Ma, He Sun, Yong Wang, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Shuai Zhou, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Qian Zhu, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692v1*, 2026.

Sicheng Xie, Lingchen Meng, Zijie Diao, Haidong Cao, Zhiying Du, Shuyuan Tu, Jiaqi Leng, Qiuyue Wang, Mingsheng Li, Shuai Bai, Zuxuan Wu, and Yu-Gang Jiang. Unify robot actions in camera frame, 2026. URL https://arxiv.org/abs/2511.17001.

XLANG Lab. Introducing roboinf: A scalable data engine for generalist robot manipulation. https://xlang.ai/blog/roboinf, May 2026. Blog post.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. ABot-M0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026.

Chao Yu, Yuanqing Wang, Zhen Guo, Hao Lin, Si Xu, Hongzhi Zang, Quanlu Zhang, Yongji Wu, Chunyang Zhu, Junhao Hu, et al. Rlinf: Flexible and efficient large-scale reinforcement learning via macro-to-micro flow transformation. *arXiv preprint arXiv:2509.15965*, 2025a.

Seungjun Yu, Seonho Lee, Namho Kim, Jaeyo Shin, Junsung Park, Wonjeong Ryu, Raehyuk Jung, and Hyunjung Shim. Waymoqa: A multi-view visual question answering dataset for safety-critical reasoning in autonomous driving. *arXiv preprint arXiv:2511.20022*, 2025b.

Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models. *arXiv preprint arXiv:2601.03309*, 2026.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, Yuxin Fan, Wenjun Li, Zhibo Chen, Fei Gao, Qi Wu, Zhizheng Zhang, and He Wang. Embodied navigation foundation model, 2025a. URL https://arxiv.org/abs/2509.12129.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. *Robotics: Science and Systems*, 2025b.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2675–2685, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023.